The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line

The FANTOM Consortium and the Riken Omics Science Center

The supplementary information contains followings;

SUPPLEMENTARY METHODS

General Methods

Cell culture and RNA extraction

THP-1 cells are heterogeneous in their ability to differentiate and respond to stimuli such as PMA and LPS[1]. To increase the signal to noise ratio in our analyses, the THP-1 cell line was sub-cloned by limit dilution and one clone (#5) was selected for ability to differentiate relatively homogeneously in response to PMA as evidenced by expression of CD14 and CSF-1R quantified by qRT-PCR (**Supplementary Fig. 1**). These THP-1 cells were frozen in aliquots, and used fresh for all subsequent experiments. All reference to THP-1 cells refer to the cloned line. THP-1 cells were cultured in RPMI, 10% FBS, Penicillin/Streptomycin, 10mM HEPES, 1mM Sodium Pyruvate and 50uM 2-Mercaptoethanol. THP-1 was treated with 30ng/ml PMA (Sigma) over a time-course of 96h. Total cell lysates were harvested in TRIzol reagent (Invitrogen) at each time-point. Undifferentiated cells were harvested in TRIzol reagent at the beginning of the PMA time-course. Total RNA was purified from TRIzol lysates according to manufacturer's instructions.

DeepCAGE

The preparation of the CAGE library from total RNA was a modification of methods described by Shiraki *et al.*[2] and Kodzius *et al.*[3], adapted to work with the 454 Life Sciences sequencer (described in detail in **Detailed Methods**).

Analysis of deepCAGE: Promoter Construction and Expression Analysis

Deep sequencing of CAGE tags was done in triplicate at 0, 1, 4, 12, 24 and 96 hours of PMA treatment for a total of 18 samples. All CAGE tags were mapped to the human genome (hg18) using the program nexalign (T. Lassmann in preparation) by aligning perfectly matching tags first, then those tags that map with a single base pair substitution and finally tags which contain a single insertion or deletion. A filter was applied to remove rRNA-derived tags. Most tags map to a unique genomic location. For tags that map to multiple locations a probabilistic model, previously described by

Faulkner *et al.*[4], was used to assign weights to each of the possible genomic mappings. The fraction of multi-mapping tags is approximately constant across samples and discarding multi-mapping tags does not affect the expression profiles across promoters (**Fig. SM-1**).

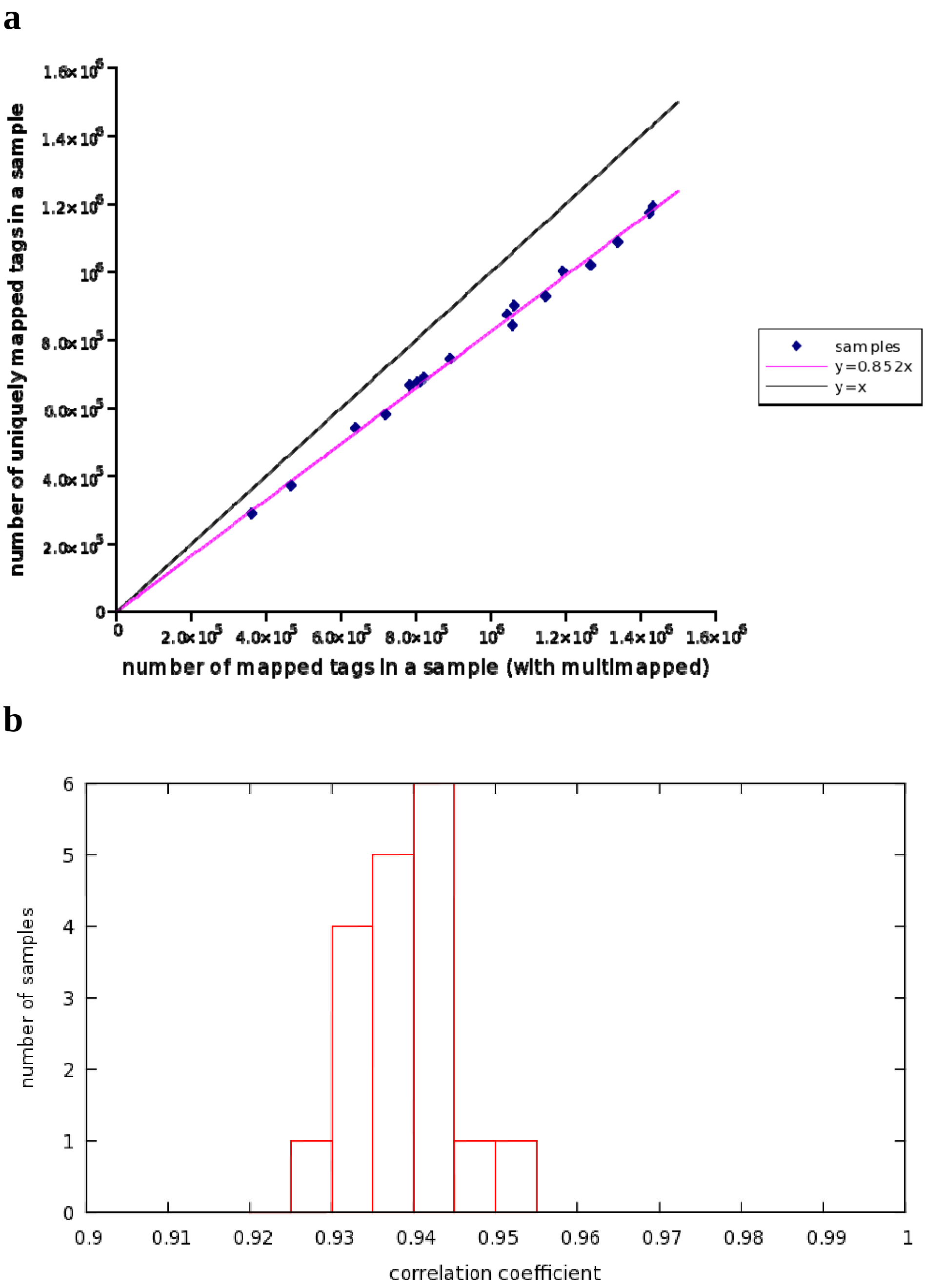


Figure SM-1 Influence of multi-mapping tags on the CAGE expression profiles. We compared CAGE expression profiles with and without all multi-mapping CAGE tags. Panel (**a**) shows that, across all samples, uniquely mapping tags are responsible for about 82.5% of all mapped tags, i.e. the large majority of tags map uniquely. Panel (**b**) shows the distribution of correlation coefficients of the expression profiles of all samples with and without multi-mapping tags. As the figure shows, there is always a very high correlation between the expression profiles with and without multi-mapping tags.

To identify promoters we first normalized the CAGE data from each sample by scaling CAGE tag counts such that the distribution of the number of tags per position matches a common reference (power-law) distribution. We used replicates to estimate experimental noise. We find that the noise distribution is well-described by a convolution of multiplicative noise and Poisson sampling noise. Using this noise model, a Bayesian procedure was used to calculate, for each consecutive pair of TSSs, the probability that both TSSs were expressed in a fixed relative proportion across all samples (**Detailed Methods**). Neighboring TSSs with a high probability of expression in a constant proportion were then hierarchically joined into clusters. Promoters were defined as significantly expressed clusters, i.e. those that have at least 1 tag in at least 2 samples and whose maximum expression across all samples is at least 10 tags per million. All other TSS clusters were discarded. Promoters within 400 bp of each other on the same strand were clustered into `promoter regions'. The distribution of the expression per TSS, promoter, and promoter region is shown in **Supplementary Figure 3**.

We obtained the genomic mappings of all human mRNAs from the UCSC BLAT alignments, discarded mRNAs whose 5' ends don't map, and then associated each promoter with all mRNAs whose mapped TSS is within 1000 base pairs of the CAGE promoter. Using the mapping from mRNAs to Entrez genes provided by NCBI we associated promoters with Entrez genes and constructed the gene locus (union of all mRNA mappings) of each gene.

We obtained phastCons conservation profiles from UCSC which are based on a multiple alignment of 28 vertebrate genomes and calculated average phastCons scores as a function of position with respect to TSS separately for promoters that are associated with genes and promoters that are not associated with any known transcript, and are more than 1000 base pairs away from any Entrez gene locus (**Supplementary Fig. 5**).

We define the normalized expression e_{ps} of promoter p in sample s as $e_{ps} = \log(t_{ps} + \tfrac{1}{2}) - \left\langle \log(t_p + \tfrac{1}{2}) \right\rangle$ where t_{ps} is the normalized number of tags per million from promoter *p* in sample *s*, and the second term is the average of the first term over the 6 time points in the replicate. For the microarray probes the expression e_{ps} is similarly given by the log-intensity of the probe in sample *s* minus the average log-intensity of the probe across the 6 time points in the replicate. Probes with detection probability less than 0.99 in all samples were discarded. For each CAGE promoter and each microarray probe we compared the total variance in the expression profile, i.e. across all time points and replicates, with the variance across replicates for each time point to estimate the fraction f_p of the total variance (FOV) that is reproducible across replicates (**Detailed Methods**). The distribution of FOV across all promoters and across all probes was summarized by their 5, 25, 50, 75, and 95 percentiles (**Supplementary Fig. 7**).

To compare deepCAGE and microarray expression measurements we associated microarray probes with CAGE promoter regions whenever the probe intersected a known mRNA whose mapped 5’ end was within 1000 bps of the promoter region. We selected all probe/promoter region pairs that are one-to-one associated with each other and calculated the Pearson correlation coefficients of their expression profiles across all samples and replicates. For each probe and promoter region we calculated an average expression profile by averaging the 3 replicate measurements at each time point, and also obtained the Pearson correlation coefficients of the average expression profiles of all probe/promoter region pairs. We collected all microarray probes that were associated with multiple CAGE promoter regions and calculated Pearson correlation coefficients between the probe expression profiles and the total expression from the associated CAGE promoter regions (summing the tags from the different promoter regions).

Binding Site Predictions

Details of the WM curation procedure are presented in the **Detailed Methods**. Briefly, we extracted all position specific weight matrices (WMs) from the JASPAR and TRANSFAC® databases that are associated with TFs of multi-cellular eukaryotes. For a few TFs (SP1[5], OCT4, NANOG[6]) we extracted WMs from the literature, and for PU.1 we inferred a new WM using the PhyloGibbs algorithm[7] (see below). WMs were associated with human TFs by matching their DNA binding domain sequences. Whenever both TRANSFAC and JASPAR WMs were available for a given TF only the JASPAR WM was used. Redundancy was removed by clustering WMs that are either highly similar themselves, are associated with equal or highly similar TFs, or predict highly overlapping sets of sites. All clusters were checked manually. For each cluster a fused WM was obtained by aligning matrices within the cluster. After a first round of prediction using these curated WMs, new matrices were constructed from the predicted sites, weighing each predicted site by its posterior probability.

For each promoter region, orthologous regions in Rhesus Macaque, Dog, Cow, Horse, Mouse and Opossum were extracted using the pairwise genome alignments provided by UCSC. The sets of orthologous sequences from 300 base pairs upstream to 100 base pairs downstream of each promoter region were aligned using T-Coffee[8]. In a completely analogous manner multiple alignments were created for the proximal promoter regions of all RefSeq starts.

TFBSs were predicted for all 201 motifs in all multiple alignments of proximal promoters using the MotEvo algorithm[9]. Like the Monkey algorithm[10] MotEvo incorporates comparative genomic information by using a specific evolutionary model for the evolution of regulatory sites for the motif as well as for the neutral background evolution. In contrast to Monkey, MotEvo incorporates the possibility that sites are under selection in only a subset of the species in the alignment. In addition, MotEvo uses a more advanced background model that distinguishes neutrally evolving background sequences from background sequences that are under purifying selection.

To incorporate the positional preferences of different motifs we adapted MotEvo to employ position-dependent prior probabilities. For each motif *m* the prior $\pi_m(x)$ denotes the probability that, in a randomly chosen promoter, a site for *m* occurs at position *x* relative to the TSS of the promoter. For each motif the prior $\pi_m(x)$ was fitted using expectation maximization starting from a uniform prior. Using the fitted priors, posterior probabilities were assigned to all predicted binding sites. Finally, all binding sites with posterior less than 0.25 were discarded and for each promoter/motif combination the score N_{pm} is given by the sum of the posterior probabilities of the remaining sites for *m* in promoter *p*. Motifs that had less than 150 predicted sites across all promoters were removed from further analysis, leaving 167 motifs.

Motif Activity Inference

With e_{ps} the expression level of promoter *p* in sample *s*, N_{pm} the predicted number of functional sites for motif *m* in promoter *p*, and A_{ms} the activity of motif *m* in sample *s*, we fit a model of the following form $e_{ps} = \text{noise} + c_p + \tilde{c}_s + \sum_m N_{pm} A_{ms}$, where c_p is a promoter-dependent constant (i.e. the basal expression of the promoter) and $\tilde{c}_s$ is a sample-dependent constant. We first fit these constants. Using the fact that $\sum_s e_{ps} = 0$ for each promoter, and defining the site-count averages $\langle N_m \rangle = \frac{1}{P} \sum_p N_{pm}$, where P is the total number of promoters, we can rewrite the model as $e_{ps} = \text{noise} + \sum_m \left(N_{pm} - \langle N_m \rangle\right) A_{ms}$. The noise is assumed to be Gaussian of unknown variance with the noise variance σ the same at each promoter (but possibly varying from sample to sample). Under this assumption the likelihood for sample *s* is given by

$$L_s \propto \sigma^{-P/2} \exp\left[-\frac{\sum_p \left(e_{ps} - \sum_m \left(N_{pm} - \langle N_m \rangle\right) A_{ms}\right)^2}{2\sigma^2} \right].$$

To minimize over-fitting we use a prior probability over activities that is centered around zero $P(A_{ms}) \propto \exp\left[-\frac{1}{2} (A_{ms}/\tau)^2 \right]$ and we set $\tau = 0.1$. The posterior distribution for the activities in sample *s* then takes the general form

$P(A_s \mid e) \propto \exp\left[-\frac{P}{\chi_s^2} \sum_{m,\tilde{m}} \left(A_{ms} - A^*_{ms}\right) C^{-1}_{m\tilde{m}} \left(A_{\tilde{m}s} - A^*_{\tilde{m}s}\right) \right]$, where the A^*_{ms} are the activities with maximal posterior probability which are determined by Singular Value

Decomposition, the activity covariance matrix $C_{m\tilde{m}}$ is a function of the site-counts N_{pm}, and χ_s^2 is the residual variance after fitting, i.e. $\chi_s^2 = \frac{1}{P}\sum_p \left(e_{ps} - \sum_m \left(N_{pm} - \langle N_m \rangle\right) A_{ms}^*\right)^2$. From this we can rigorously calculate a standard-error σ_{ms} for the activity of each motif in each sample, and calculate a z-value $z_{ms} = A_{ms}^* / \sigma_{ms}$. Note that, given the Gaussian form of the posterior for the activity of each motif, the p-value for the significance of the motif's activity can be directly determined from the z-value. Finally, we calculate an overall significance of the motif by averaging its z-value over the samples $z_m = \sqrt{\frac{1}{S}\sum_s (z_{ms})^2}$ where S is the number of samples. Analogously, the posterior distribution of motif activities is inferred from the expression profiles of microarray probes and the site-counts in associated promoters. Final motif activities A_{mt} as a function of time are inferred by combining the posterior distributions from the 3 replicates for both CAGE and the microarrays assuming one underlying activity for each motif at each time point.

To quantify the quality of the fits we first calculate the `expression signal', i.e. the total variance that could possibly be explained by the fit. The expression variance of a promoter is given by $v_p = \frac{1}{S}\sum_s (e_{ps})^2$ and with f_p the FOV for this promoter, the total expression signal is $E = \sum_p f_p v_p$. The fraction ρ of expression variance explained by the fit is then $\rho = \frac{\sum_s \left(\sum_p (e_{ps})^2 - \chi_s^2\right)}{E}$.

To select core motifs we combined the posterior distributions over motif activities from the posterior distributions of the 3 replicates for both CAGE and the microarrays (**Detailed Methods**). The result is a final average motif activity A_{mt}^f for each motif at each time point, plus a standard-error σ_{mt}^f. Using this we calculate a final significance z_m^f for each motif. In addition we calculate the fraction of variance in motif activity that is reproduced across the replicates of both CAGE and microarray (FOV, **Detailed Methods**). The 30 selected core motifs are all motifs with z-values at least 3.75 and FOV at least 0.75 (**Fig. 2**).

We clustered the activity profiles of the core motifs using a Bayesian hierarchical clustering method (**Detailed Methods**). Briefly, starting from the posterior distributions of motif activities for all motifs we can calculate, for any pair of motifs, the probability that their activity profiles are the same (i.e. within noise). We iteratively clustered the two motifs with highest probability of being the same and determined the new posterior probability of motif activities for the cluster. We stopped when the probability for the highest scoring pair fell below a cut-off that we determined by hand.

Motif target predictions

We predict a regulatory edge between a motif and a promoter when the promoter has predicted binding sites for the motif ($N_{pm} \geq 0.25$) and the expression profile of the promoter correlates significantly with the inferred final activity profile A^f_{mt} of the motif. In particular, the correlation between the expression profile and activity profile is given by $c_{pm} = \frac{1}{6}\sum_t e_{pt} A^f_{mt}$, where e_{pt} is the time-dependent expression profile of the promoter averaged over the replicates. Using only a single motif to explain the expression profile, the residual variance is $\chi^2_{pm} = \frac{1}{6}\sum_t \left(e_{pt} - c_{pm} A^f_{mt}\right)^2$. Finally, the z-value that quantifies the significance of the regulatory interaction between motif and promoter is $z_{pm} = \sqrt{\frac{6}{\chi^2_{pm}}} c_{pm}$.

Note that, although c_{pm} can be negative, we only consider regulatory interactions with non-negative correlation. For the Gene Ontology analysis, target gene sets of core motifs (z-value ≥ 1.5 for the association of a motif to promoters of target genes, z-values were averaged if there was more than one promoter associated with a gene) were tested for functional enrichment[11]. All genes with CAGE defined promoters were chosen as the background.

siRNA edge validation and core network construction

Predicted regulatory interactions were tested using siRNA knockdowns of 28 TFs that are associated with motifs. For each TF knockdown we collected all microarray probes that are associated with promoters and calculated, for each probe, the average z-value of the predicted regulatory interaction from the TF's motif to the promoters associated with the probe. At different cut-offs in z-value we then divided the probes into `targets' of the motif, i.e. those with a z-value above the cut-off, and `non-targets' of the motif, i.e. all probes with z-value below the cut-off (this includes probes for which there are no predicted TFBSs in the associated promoters), and calculated the difference in average expression ratio (knockdown minus mock) of targets and non-targets. For each knockdown we calculated the Pearson correlation coefficient between the z-value cut-off on target prediction and the observed difference in average expression ratio of targets and non-targets. To assess the significance of the differences in average expression ratio we set an intermediate cut-off of z=1.5, calculated the distribution of expression ratio for targets and non-targets, determined their means (μ_t and μ_{nt}) and variances (v_t and v_{nt}), and determined a z-value for the expression ratio difference as

$z = \frac{\mu_t - \mu_{nt}}{\sqrt{v_t / N_t + v_{nt} / N_{nt}}}$, where N_t and N_{nt} are the number of target and non-target probes, respectively.

The core network was constructed by first selecting all predicted regulatory interactions (z-value at least 1.5) between core motifs and promoters that are associated with a gene which is a TF that in turn is associated with a core motif. This set of predicted regulatory interactions was then filtered by choosing only interactions that have independent experimental support of at least one of the following types. 1) The regulatory interaction has been reported in the literature 2) There is a ChIP-chip experiment in which binding of one of the TFs associated with the motif to the promoter of the target gene has been reported. 3) In our siRNA experiments the target promoter is observed to be perturbed in expression (B-statistic larger than zero) after knockdown of a TF associated with the motif.

Motif Activity Analysis of TF knockdowns

We applied the motif activity analysis to the microarray expression profiles of all siRNA samples including negative controls. As a result we obtained fitted motif activities A^{*}_{ms} and standard-errors σ_{ms} for each motif *m* in each of the siRNA samples *s*. We combined the inferred activities from replicates and control experiments, and calculated a z-value for the activity change between siRNA and negative control for each TF that was knocked down:

$z^{TF}_{m} = \dfrac{\left\langle A^{TF}_{m} \right\rangle - \left\langle A^{NC}_{m} \right\rangle}{\sqrt{(\sigma^{TF}_{m})^2 + (\sigma^{NC}_{m})^2}}$, where $\left\langle A^{TF}_{m} \right\rangle$ is the average activity of motif *m* across the replicates in which the TF was knocked down, σ^{TF}_{m} the standard-error of this average activity, $\left\langle A^{NC}_{m} \right\rangle$ the average activity of motif *m* in the negative controls, and σ^{NC}_{m} its the standard-error. The z-values z^{TF}_{m} characterize the expression changes observed upon siRNA knockdown of the TF in terms of observed changes in motif activities. That is, if z^{TF}_{m} is highly positive it indicates that predicted targets of motif *m* are up-regulated in response to knockdown of the TF. We similarly calculated z-values for motif activity changes across the PMA time course:

$z^{PMA}_{m} = \dfrac{\left\langle A^{96}_{m} \right\rangle - \left\langle A^{0}_{m} \right\rangle}{\sqrt{(\sigma^{96}_{m})^2 + (\sigma^{0}_{m})^2}}$, where $\left\langle A^{96}_{m} \right\rangle$ is the average activity of motif *m* after 96 hours of PMA treatment, σ^{96}_{m} its standard-error, $\left\langle A^{0}_{m} \right\rangle$ is the average activity before PMA treatment, and σ^{0}_{m} its standard-error. Given the z-value for the change in motif activity the probability that the motif is up-regulated is given by $p_{\mathrm{up}}(z) = \frac{1}{2}\mathrm{Erfc}\left(\frac{z}{\sqrt{2}}\right)$ and the probability that the motif is down-regulated is given by $p_{\mathrm{down}}(z) = \frac{1}{2}\mathrm{Erf}\left(\frac{z}{\sqrt{2}}\right)$. Using this we calculated, for each motif *m*, the probability p_m that the motif is changing in the same direction in both the PMA time course and the TF knockdown:

$$p_m = p_{\text{up}}(z_m^{TF})p_{\text{up}}(z_m^{PMA}) + p_{\text{down}}(z_m^{TF})p_{\text{down}}(z_m^{PMA}) .$$

Finally, the overlap o^{TF} between TF and PMA time course is defined as the sum of p_m over all motifs divided by the total number of motifs, i.e. the estimated fraction of motifs that change activity in the same direction in knockdown and PMA time course. We calculated the significance of the differentiative overlaps by a permutation test; we randomly permuted the order of the motifs 1000 times and calculated the differentiative overlap for each.

Illumina microarray analysis

THP-1 samples were identical to those used for deepCAGE libraries, and RNA was purified for expression analysis by Qiagen RNeasy columns, Takara FastPure RNA Kit or TRIzol. RNA quality was checked by Nanodrop and Bioanalyser. RNA (500 ng) was amplified using the Illumina TotalPrep RNA Amplification Kit, according to manufacturer’s instructions. cRNA was hybridized to Illumina Human Sentrix-6 bead chips Ver.2, according to standard Illumina protocols (http://www.illumina.com). Chips scans were processed using Illumina BeadScan and BeadStudio software packages and summarized data was generated in BeadStudio (version 3.1). Quantile normalization of Illumina data and B-statistic calculations were carried out using the lumi and limma packages of Bioconductor in the R statistical language[12-14]. For differential gene expression during the timecourse and between siRNAs and negative control transfections we required a B-statistic ≥ 2.5, expression ratio ≥ 2 and the gene had to be detected in one of the conditions (average detection score ≤ 0.01).

Transcription factor expression classes and their regulatory inputs

Expressed TFs were defined from a previously published list of curated human TFs[15] and required detection by both CAGE and Illumina in at least one timepoint (and in all 3 biological replicates). For Illumina microarray the average detection score had to be less than 0.01. For CAGE the threshold was ≥ 3TPM (tags per million). Dynamic expression was then determined using the Illumina array data and B-statistics (see above). Undifferentiated was defined by contrasting 0h vs 96h with an expression ratio < -2, Differentiated was defined by contrasting 0h vs 96h and expression ratio > 2.

Transient dynamic genes were defined by 0h vs the remaining timepoints (1h, 4h, 12h, 24h). The 1hr up-regulated set was defined by only the subset of 0h vs 1h and expression ratio >2. Edge prediction enrichment in these classes was determined by extracting the set of predictions for all 610 detected TFs, and then comparing the number of predictions in class versus the remaining detected factors. P-value was calculated using Fisher's exact test.

ChIP on chip analysis

Details of the analysis are described in the **Detailed Methods**. Briefly, THP-1 cells were cross-linked with 1% formaldehyde for 10 min and cells were collected by centrifugation and washed twice in cold 1 x PBS. The cells were sonicated for 5~7 min with a Branson 450 Sonicator to shear the chromatin. Complexes containing DNA were immunoprecipitated with antibodies against H3K9Ac (07-352, Upstate), PU.1 (T-21, Santa-cruz), SP1 (07-645, Upstate), and RNA Polymerase II (8WG16, Abcam), respectively. The immunoprecipitated sample was incubated with magnetic beads/Protein G (Dynal) for 1 hr at 4°C followed by washing. The complexes were eluted from the magnetic beads by addition of 1% SDS and 100 mM $NaHCO_3$. Beads were vortexed for 60 min at RT. The supernatants were incubated for 3.5 hr at 65°C to reverse the cross-links, and incubated with RNaseA, and then proteinase K, followed by a phenol:chloroform:isoamyl alcohol extraction and ethanol precipitation to recover the DNA. Immunoprecipitated DNA was amplified by either linker-mediated PCR (LM-PCR) or *in vitro* transcription (IVT) followed by synthesis of double-strand cDNA. Amplified DNA was end-labeled with biotin-ddATP and was hybridized to Affymetrix whole genome tiling or promoter arrays.

A new PU.1 DNA binding motif was inferred by extracting the top 50 bound genomic segments, extracting orthologous regions from Rhesus Macaque, mouse, cow, and dog, multiply aligning them, and running the PhyloGibbs algorithm[7] on these alignments. We observed that the resulting motif better distinguishes bound from unbound regions genome-wide than the PU.1 matrix from TRANSFAC (data not shown).

Perturbation experiments

THP-1 cells were seeded in 6 cm dishes at a density of 1 × 10^6 cells/dish for transfection. Transfection was performed with 1.6 μg/ml (final concentration) of Lipofectamine 2000 (Invitrogen) and 20 μM (final concentration) of stealth siRNA (Invitrogen) or negative control (Stealth™ RNAi Negative Control Medium GC Duplex – catalog number 12935-300, Invitrogen) by reverse transfection protocol in accordance with the manufacturer's instructions. Total RNA for Illumina microarray analysis was extracted 48h after transfection, using the FastPure RNA kit (TAKARA BIO, Ohtsu, Shiga, Japan) in accordance with the manufacturer's instructions. TF gene expression levels in THP-1 cells treated with gene-specific siRNAs or the calibrator negative control siRNA were estimated by qRT-PCR in triplicate. Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) mRNA levels were determined concurrently, for normalization. All microarray experiments were conducted in biological triplicate, and the effects of each TF knockdown was assessed relative to the duplex negative control transfection. The sequences of siRNAs and qRT-PCR primers used in the present study are shown in **Supplementary Table 7**.

FACS analysis of THP-1 knockdown cells

All FACS experiments were performed in triplicates wells processed individually. In wells with both suspended and adherent THP-1 cells, suspended cells were harvested and adherent cells then detached by two incubations with 0.5mM EDTA for 2 min each after which they were pooled with suspended cells from the same well. Cells were washed once with flow wash buffer (0.1% BSA in PBS) and resuspended in flow wash buffer. THP-1 cells were stained with 2μl antibody for 20 min at 4°C in 100 *μl* flow wash buffer, washed 1x with 1ml flow wash buffer and resuspended in 250μl flow wash buffer for analysis. Data acquisition was performed on a BD FACSCanto II with BD FACSDiva software used for analysis. The following antibodies were used: CD9-FITC, CD11b-PE, CD14-PE, CD15-FITC, CD18-FITC, CD54-PE, CD105-FITC and MPO-FITC from Immunotools, and CD11c-PE, CD192-Alexa647 (CCR2) and HLA-DR-PECy5 from Becton-Dickenson.

The data and analysis results available from the FANTOM4 web resource

In addition to the data and analysis results amassed here the full set of several **Supplementary Tables** are available from the FANTOM4 web resource. The data is also available from "GNP Platform" (http://genomenetwork.nig.ac.jp/index_e.html).

Detailed Methods

1.0 DeepCAGE

1.1 Preparation of CAGE libraries.

1.1.1 Synthesis of first-strand cDNA

cDNA was synthesized using 50 μg total RNA in 20 μl water and 2 μl random primer (N20; 6 $\mu g / \mu l$) with M-MLV Reverse Transcriptase RNase H Minus, Point Mutant (Promega). The RNA and primer were heated to 65 $^\circ C$ for 5 min and then placed on ice. The reaction mixture, 75 μl of 2x GC I LA Taq buffer (TaKaRa), 4 μl of 10mM dNTPs, 30 μl of Sorbitol/Trehalose mix, 4 μl of water and 15 μl of Reverse Transcriptase (200 $U / \mu l$) was then added, followed by a reverse transcription step in a thermal cycler as follows: 30 s at 25 $^\circ C$, 30 min at 42 $^\circ C$, 10 min at 50 $^\circ C$, 10 min at 56 $^\circ C$. The reaction was stopped with EDTA and proteinaseK was added. cDNA/RNA hybrids were purified by CTAB precipitation and the pellet was completely dissolved in 46 μl of water.

1.1.2 Oxidation/biotinilation

3.3 μl 1M Sodium Acetate (pH 4.5) and 2 μl 250mM $NaIO_4$ were mixed and incubated for 45 min on ice, in the dark. The reaction was stopped with glycerol, and cDNA/RNA hybrids were precipitated with isopropanol. The pellet was dissolved in 50 μl water and 5 μl 1M Sodium Citrate (pH6.1), 5 μl 10% SDS and 150 μl 10mM Biotin (long arm) Hydrazide was added. The reaction was incubated for 10-12h at room temperature. The reaction was then stopped with 75 μl 1M Sodium Acetate (pH 6.1) and precipitated with ethanol. The pellet was dissolved in 180 μl 0.1xTE and treated with RNaseI (1U/ µg starting RNA). The sample was then digested with proteinaseK, purified with phenol/chloroform and cDNA/RNA hybrids were precipitated with isopropanol. The pellet was resuspended in 50 μl 0.1xTE.

1.1.3 Capture-release

500 μl dynabeads MP-280 streptavidin (Dynal) were blocked with 100 μg tRNA for 30 min on ice with occasional shaking. The beads were washed 3 times with 500 μl wash buffer (4.5M NaCl / 50mM EDTA) and resuspended in 500 μl wash buffer. The beads were added to cDNA/RNA hybrids and incubated with mild agitation at 50 $^\circ C$

for 30 min to allow the biotinylated cap to bind to the beads. The cap/bead complexes were collected using a magnetic stand. The beads were sequentially washed with the following buffers: 2 further washes with wash buffer, 1 wash with 0.3M NaCl / 1mM EDTA, 3 washes with 0.4% SDS / 0.5M NaOAc / 20mM Tris-HCl pH8.5 / 1mM EDTA and 2 washes with 0.5M NaOAc / 10mM Tris-HCl pH8.5 / 1mM EDTA. The captured full-length cDNAs were eluted from the beads using 3 washes with 100 μl elution buffer (50mM NaOH / 5mM EDTA). The eluted cDNAs were immediately transferred to ice and 100 μl 1M Tris-HCl (pH 7.0), 10 units of RNaseI was added and the mixture was incubated at 37 °C for 10 min. Samples were digested with proteinaseK, purified with phenol/chloroform and cDNAs were precipitated with isopropanol. The sample was further purified using a S400 MicroSpin column, isopropanol precipitated and resuspended in 5 μl water.

1.1.4 Single-strand linker ligation

Specific double-stranded linkers were used to concatenate deepCAGE tags (**Table SM1**). These contained a recognition site for *Xma*JI, a 5bp tissue identification tag (XXXXX, that allowed the tag to be assigned to its original RNA source during construction of a pooled CAGE library) and the class II restriction enzyme *Mme*I. Upper oligonucleotides, GN5 and N6, were mixed in a 4:1 ratio, before this solution was mixed in a 1:1 ratio with the lower oligonucleotide to give a double stranded linker with a (G/N)NNNNN overhang after annealing. 0.2 μg linker was added to the single-strand cDNA prepared above (**section 1.1.3**). Using TaKaRa Ligation Kit ver.2.1, the sscDNA:linker mixture was ligated at 16 °C overnight. Following ligation, the abundance of beta-actin cDNA (*ACTB*) in each library was quantified by qRT-PCR so that equivalent amounts of each library could be pooled to construct a mixed CAGE library. ProteinaseK and phenol/chloroform extraction was used to remove enzymes, and cDNA was purified away from free linker using an S400 spin column. cDNA was ethanol precipitated and dissolved in 10 μl 0.1x TE.

1.1.5 Second-strand synthesis

In order to synthesize the complementary cDNA strand, 6 μl second-strand primer (**Table SM1**) at 100 ng / μl, 7.2 μl of 5x buffer A, 4.8 μl of 5x buffer B (Invitrogen), 6 μl 2.5mM dNTPs were added to 10 μl cDNA sample, and water was added to a final volume of 58μl. The reaction mixture was heated to 65$^\circ C$, and 2μl Elongase polymerase (1$U/\mu l$; Invitrogen) was added. The reaction was incubated in a thermalcycler at 5 min/65$^\circ C$, 30 min/68$^\circ C$ and 10 min/72$^\circ C$. cDNA was purified away from free oligonucleotides using a S400 spin column, ethanol precipitated and dissolved in 10μl 0.1x TE.

Table SM1 Linker oligonucleotides used for deepCAGE tag concatenation.

Linker oligo name	Sequence (5’-3’)
upper oligonucleotide GN5	biotin-agagagagacctcgagtaactataacggtcctaaggtagcgacctaggXXXXXtccgacGNNNNN
upper oligonucleotide N6	biotin- agagagagacctcgagtaactataacggtcctaaggtagcgacctaggXXXXXtccgacNNNNNN
lower oligonucleotide	Pi-gtcggaXXXXXcctaggtcgctaccttaggaccgttatagttactcgaggtctctctct-NH2
second-strand primer	biotin-agagagagacctcgagtaactataa

1.1.6 Tagging

The double-stranded linker:cDNA complex was cleaved with the class II restriction enzyme, *Mme*I (3 units/µg cDNA) in a 100μl reaction incubated at 37$^\circ C$ for 1h. Following purification as above (ProteinaseK, phenol/chloroform, ethanol), 1.6μg the 2nd linker (**Table SM2**) was added to the sample in a total reaction volume of 20μl, heated to 65$^\circ C$ for 2 min, and then set on ice. The 2nd linker was ligated to the captured cDNA with T4 DNA ligase at 16$^\circ C$ overnight. The reaction was stopped by heating to 65$^\circ C$ for 5 min and 80μl 0.1xTE buffer was added. 200μl Dynabeads M-280 streptavidin beads were blocked with 40μg tRNA for 30 min on ice with occasional shaking. The beads were washed 3 times with 200μl 1x B+W buffer (1M NaCl / 0.5mM EDTA / 5mM Tris-HCl pH 7.5) and resuspended in 100μl 2x B+W buffer. Streptavidin beads were mixed with biotinylated CAGE tags and incubated with mild agitation at room temperature for 15 min to allow binding. The CAGE tag/bead complexes were then collected with a magnetic stand. The beads were washed sequentially: twice with 200μl 1x B+W buffer containing BSA (200$\mu g/ml$), twice with 200μl 1x B+W buffer and twice with 200μl 0.1xTE buffer. The 5’-end cDNA tags were released from the beads by adding excess free biotin. The elution was repeated three times, and fractions were pooled. 3.5μg glycogen was added, the sample was ethanol precipitated and resuspended in 50μl 0.1x TE buffer. The cDNA tags were treated with *RNa*seI and further purified with a G50 spin column and ethanol precipitated. The sample was resuspended in 24 μl water.

Table SM2 Linker oligonucleotides used for deepCAGE tag concatenation.

Linker oligo name	Sequence (5’-3’)
2nd linker forward	Pi-cctaggtcaggactcttctatagtgtcacctaaagacacacacac-NH2
2nd linker reverse	gtgtgtgtgtctttaggtgacactatagaagagtcctgacctaggNN

1.1.7 Amplification of CAGE tags

DNA fragments were amplified by PCR using the following two linker-specific primers: Primer1: 5’-biotin-CTATAGAAGAGTCCTGACCTAGG-3’; Primer2: 5’-biotin-CGGTCCTAAGGTAGCGACCTAG-3’. Ten parallel PCRs were performed in a total volume of 50 μl, using 1.6 μl cDNA-tags / 5 μl 10x PCR buffer / 3 μl DMSO / 12 μl 2.5mM dNTPs / 0.5 μl Primer1 (350 ng / μl) / 0.5 μl Primer2 (350 ng / μl) / 26.6 μl water / 0.8 μl DNA polymerase (2.5 U / μl). Samples were incubated at 94 °C for 1 min, and 20 cycles of 30 sec/94 °C, 20 sec/55 °C and 20 sec/70 °C were performed, followed by a 5 min incubation at 72 °C. The resulting PCR products were pooled, purified, isopropanol precipitated and resuspended in 24 μl 0.1xTE buffer.

PCR products were further purified by polyacrylamide gel electrophoresis. The 75 bp band was cut out of the gel, crushed, and incubated with 150 μl in buffer (2.5mM Tris-HCl pH7.5 / 1.25M ammonium acetate / 0.17mM EDTA pH7.5) overnight at room temperature. The extracted tags were filtered using MicroSpin columns. A further 150 μl buffer was added to the remaining gel, and rotated at room temperature for 30 min. This extraction step was repeated a further 3 times. The extracted tags were precipitated with ethanol and dissolved in 30 μl 0.1xTE. The DNA concentration was measured with Picogreen.

Purified bands were again PCR-amplified in a total of 100 μl in a reaction containing 0.2-1ng of cDNA-tags/10 μl of 10x PCR buffer/6 μl of DMSO/12 μl of 2.5mM dNTPs/0.75 μl of Primer1 (1 μg / μl)/ 0.75 μl of Primer2 (1 μg / μl)/ 0.8 μl of DNA Polymerase (2.5 U / μl). Samples were heated to 94 °C for 1 min, and subjected to 8 cycles of 30 sec/ 94 °C, 20 sec/55 °C and 20 sec/70 °C, followed by a final elongation at 72 °C for 5 min. PCR products were pooled, purified, ethanol precipitated and finally redissolved in 50 μl of 0.1xTE. To eliminate excess primers, PCR products were further purified using MinElute columns (Qiagen), ethanol precipitated and redissolved in 100 μl of 0.1x TE. DNA concentration was again measured with Picogreen.

Purified PCR products were digested with *Xma*JI (2µg/tube), followed by ProteinaseK treatment in 200 μl . The desired 37 bp DNA tags were separated from unwanted DNA using streptavidin-coated magnetic beads, which retained the biotin-labeled DNA. The cleaved tags were mixed with 500 μl beads and incubated at room temperature for 15 min with mild agitation to allow binding. The magnetic beads were removed and the 37 bp tags were extracted by phenol/chloroform followed by ethanol precipitation and dissolved in 45 μl of TE.

The tags were further purified using polyacrylamide gel electrophoresis as above. Purified tags were resuspended in 6 μl of 0.1xTE, and DNA quantitated with picogreen.

500ng CAGE tags were ligated to form concatemers by addition of 6 μl tag DNA to 1.0 μl 10x T4 DNA ligase buffer, 1.0 μl T4 DNA ligase and 454 adaptors A/B as described in the original publication[16], in a reaction of 10 μl incubated overnight at 16 $^\circ C$, and treated with ProteinaseK. The sample was purified with a GFX column to eliminate short concatemers. The eluted sample was transferred for sequencing.

1.2 CAGE tag sequencing

1.2.1 Single-stranded template DNA (sstDNA) preparation

Concatenated CAGE tags were immobilized with the pre-washed immobilization beads (GS20 DNA Library Preparation Kit) via biotin:streptavidin interactions during a 20 minute incubation. The immobilized beads were washed with the GS20 DNA Library Wash Buffer to remove biotin-unlabelled dsDNAs. sstDNAs were recovered from the immobilized dsDNAs using 0.125N NaOH. The sstDNAs were neutralized with acetic acid and were purified with a MinElute PCR purification kit (Qiagen). The amount and the average length of sstDNAs were measured by Agilent 2100 BioAnalyzer with a RNA Pico 6000 LabChip to estimate the concentration of sstDNAs.

1.2.2 Emulsion PCR for titration assay

To determine optimal amplification conditions, sstDNAs and pre-washed Capture Beads for amplification (GS20 emPCR kit) were mixed together in a range of ratios. The sstDNAs were annealed to the capture beads using the thermocycler sstDNA annealing program according to the manufacturer's protocol. The emulsion was independently made with Emulsion Oil and Mock Amplification Mix by shaking TissueLyser (Qiagen), following the manufacturer's protocol. After the emulsification step, the sstDNA-annealed Capture Beads and the Live Amplification Mix containing Amplification Mix, MgSO4, Amplification Primer Mix, Platinum HiFi Taq Polymerase, and PPiase, was added to the emulsion tube. Another shaking step created an emulsion with aqueous phase micelles of the appropriate size to contain single beads with amplification mix. The resulting emulsion reaction was then subjected to PCR amplification, and the Capture Beads were recovered following the manufacturer's protocol.

1.2.3 Emulsion PCR for sequencing in large scale

Once the optimal amplification conditions, especially the optimal mixing ratio, were determined by titration, sstDNAs and the Capture Beads were mixed in optimal proportions for large-scale sequencing. The annealing, emulsification, PCR and bead recovery steps were performed as described above.

1.2.4 Sequencing

The Capture Beads on which DNAs were successfully amplified were enriched using the GS20 enrichment beads prior to sequencing. GS20 sequencing primers were annealed to the enriched sstDNA beads ("DNA-carrying beads"), using the sequencing primer annealing program specified by the manufacturer. After the completion of the primer annealing procedure, the number of enriched beads was counted with a Coulter Counter (Beckman Coulter). The appropriate number of beads was applied to a small or large PicoTiterPlate with packing beads and enzyme beads (also from GS20 sequencing kit). The sequencing run and the base call analysis were performed following the manufacturer's protocol.

2.0 Analysis of deepCAGE

2.1 CAGE tag extraction

Each unit in a concatamer consists of the following ordered sub-units: a-b-c-t-a, where a = *Xma*JI restriction site, b = barcode sequence indicating time point, c = *Mme*I recognition site for producing the CAGE tag, t = tag sequence. During tag extraction, sequences containing one or more undetermined bases ("N") were discarded from further analysis. An in-house analysis program was used to find instances of a-b-c-t-a and the reverse complement counterpart a'-t'-c'-b'-a'. Only units containing perfect matches to *Xma*JI, *Mme*I and barcode sub-units, and tag lengths of between (and including) 18-24 bp, were extracted.

2.2 CAGE tag mapping

A novel alignment method, nexalign, was used to align all CAGE tags to the human genome reference sequence (hg18) using a layered, iterative approach. Firstly, tags were matched exactly to the genome and their positions recorded. Secondly, tags that did not match in the first pass were subjected to single base pair substitutions at every position and realigned. Finally, those tags that still did not map were subjected to mapping with indels and aligned to the genome. After this, the match that contained the fewest errors for a given tag was designated the "best" match.

For the majority of tags the "best" match was unique on the genome. However, if a tag matched multiple locations at a best match level, a multi-mapping CAGE tag rescue strategy, previously described by Faulkner *et al.*[17] was used to assign tags to their most probable location. Finally, a filter was applied to remove rRNA-derived tags.

2.3 CAGE expression normalization, noise analysis, and promoter construction

The detailed procedures and mathematical derivations involved in our normalization, noise-analysis, and promoter construction are beyond the scope of the current manuscript and will be described elsewhere (Balwierz and van Nimwegen, submitted). Here we provide a summary of the crucial steps.

First, we observed that for each of our 18 deepCAGE samples (6 time points in triplicate) the number of different TSS positions observed is very large (1.85 million when combining data from the 18 samples), but the large majority of TSS positions have a very small number of tags. In particular, we find that in each sample, the distribution of the number of tags per TSS is power-law distributed. We fitted power-laws to the tags-per-TSS distribution for each sample and found that the exponent of the fitted power-law varied between -1.3 and -1.2. The off-sets varied by a factor of about 3 across the 18 samples, reflecting the variance in total number of tags that were mapped. To normalize the deepCAGE expression data we chose a reference power-law distribution with exponent -1.25 and an off-set corresponding to a total of 1 million tags. We then transformed the tag counts from each TSS in each sample such that, after transformation, all samples obey the same reference distribution of tags per TSS. Specifically, for each sample parameters α and β are chosen, and all raw tag counts are transformed according to $t \rightarrow t' = \alpha t^{\beta}$ such that the distribution of t' matches the reference distribution.

Second, we find that the distribution of noise in CAGE expression can be well-fitted by a convolution of multiplicative (log-normal) noise, and Poisson sampling noise. The final form of the noise distribution we find is as follows. If x is the logarithm of the true frequency of a given TSS in the pool of TSSs, then the probability to measure n tags, corresponding to a log-frequency of $y = \log\left(n/N\right)$ among the set of N mapped tags is approximately given by

$$P(y,n \mid x,\sigma,N) \approx C\exp\left(-\frac{1}{2}\frac{(x-y)^2}{\sigma^2 + 1/n}\right)$$

where C is a normalization constant and σ is the size of the multiplicative noise. The latter is estimated from replicate deepCAGE data-sets to be $\sigma \approx 0.245$.

Using this noise model we can calculate, for any pair (t_1, t_2) of neighboring TSSs on the genome, the probability $P(t_1, t_2)$ of their observed expression profiles (tag counts in all 18 samples) under the assumption that the ratio of true expression of the TSSs is *constant* across all samples, i.e. the expression profiles are directly proportional to each other. Similarly we can calculate the probability $P(t_1)P(t_2)$ of the two expression profiles assuming they are independent. We use these probabilities to hierarchically cluster neighboring TSSs into `promoters'. For each pair of TSSs that are a distance d apart we assign a prior probability $\pi(d) = \exp(-d/10)$ that they belong to a common promoter. The posterior probability that the pair (t_1, t_2) belongs to a common promoter is then $\frac{P(t_1,t_2)\pi(d)}{P(t_1,t_2)\pi(d) + P(t_1)P(t_2)(1-\pi(d))}$. We iteratively cluster pairs of neighboring promoters with highest posterior probability until the highest posterior probability is less than ½. Finally, to choose significantly expressed promoters we retained only those promoters that have at least 1 tag in at least 2 samples and at least 10 tags-per-million in at least 1 sample. For each remaining promoter we defined the proximal promoter as the segment from 300 bps upstream of the first TSS in the promoter to 100 bps downstream of the last TSS in the promoter. Promoters with overlapping proximal promoters on the same strand were clustered into *promoter regions*. Thus, our analysis identifies promoters hierarchically at three levels: The 'TSS level' of individual transcription start sites, the 'promoter level' of clusters of nearby TSSs with indistinguishable expression profiles, and the 'promoter region' level of promoters with overlapping proximal promoters.

2.4 Expression signal versus replicate noise

For each promoter we estimated the fraction of the variance in its expression values that could be explained theoretically, i.e. the fraction that is not due to noise. To do this we compared the variance of expression at the same time point across replicates with the total variance, i.e. across all replicates and time points. For each promoter *p* we started from the log-expression values $x_s^i = \log\left[t_s^i + \tfrac{1}{2}\right] - \left\langle \log\left[t^i + \tfrac{1}{2}\right]\right\rangle$, where t_s^i is the normalized tag-per-million count of the promoter in replicate *i* and time point *s*, and the average in the second term is over the 6 time points in the replicate. That is, $\sum_s x_s^i = 0$

for each replicate *i* when summed over the time points *s*. We assume that x_s^i is the sum of a `true' expression value δ_s (which is of course the same for all replicates) and replicate noise. We denote by σ^2 the size of the replicate noise, and by τ^2 the size of the variance in true expression. Using this the prior probability of the true expression values is given by a Gaussian:

$P(\delta_s \mid \alpha) = \sqrt{\frac{\alpha}{2\pi}} \exp\left(-\frac{\alpha}{2} (\delta_s)^2 \right)$ where $\alpha = \frac{1}{\tau^2}$. Similarly the probability of the observed expression values given the true expression values and size of the noise is

$P(x_s^i \mid \delta_s, \beta) = \sqrt{\frac{\beta}{2\pi}} \exp\left(-\frac{\beta}{2} \left(x_s^i - \delta_s\right)^2 \right)$ where $\beta = \frac{1}{\sigma^2}$. Using these two expressions we obtain for the probability of the data given α and β

$$P(x_s \mid \alpha, \beta) = \int_{-\infty}^{\infty} P(\delta_s \mid \alpha) \prod_{i=1}^{r} P(x_s^i \mid \delta_s, \beta) d\delta_s \propto \sqrt{\frac{\alpha}{\alpha + \beta r}} \beta^{r/2} \exp\left(-\frac{1}{2} \left[\frac{\beta^2 r^2}{\alpha + \beta r} \operatorname{var}(x_s) + \frac{\alpha\beta r}{a + \beta r} \left\langle (x_s)^2 \right\rangle \right] \right)$$

where *r* is the number of replicates (3 in our case), $\mathrm{var}(x_s) = \frac{1}{r}\sum_{i=1}^{r}\left(x_s^i - \langle x_s \rangle\right)^2$ is the variance in expression across the replicates for time point *s*, and $\left\langle (x_s)^2 \right\rangle = \frac{1}{r}\sum_{i=1}^{r}\left(x_s^i\right)^2$ is the average squared log-expression at time point *s*. To get the probability over all time points we simply take the product of the above expression over all time points, i.e. $P(x \mid \alpha,\beta) = \prod_s P(x_s \mid \alpha,\beta)$. We are interested in calculating the fraction *f* of the total expression variance (FOV) that is reproducible across the replicates. This fraction *f* is given by $f = \frac{\tau^2}{\sigma^2+\tau^2} = \frac{\beta}{\alpha+\beta}$. We write $P(x \mid \alpha,\beta)$ in terms of *f* and β and we integrate over β to obtain the probability of the data as a function of *f* only, i.e. $P(x \mid f) = \int P(x \mid f,\beta)\frac{d\beta}{\beta}$. We then finally find:

$$P(x \mid f) \propto \left(\frac{1-f}{1+(r-1)f}\right)^{n/2}\left(\frac{rf\mathrm{var}(x)+(1-f)\left\langle (x)^2 \right\rangle}{1+(r-1)f}\right)^{-nr/2},$$

where *n* is the number of time points (6 for our case), $\mathrm{var}(x) = \frac{1}{n}\sum_{s=1}^{n}\mathrm{var}(x_s)$ is the variance across replicates averaged over all time points, and $\left\langle (x)^2 \right\rangle = \frac{1}{n}\sum_{s=1}^{n}\left\langle (x_s)^2 \right\rangle$ is the average squared log-expression across all replicates and time points. Finally, we use the expression $P(x \mid f)$ to calculate the expected value of *f*, i.e. $\langle f \rangle = \frac{\int fP(x \mid f)df}{\int P(x \mid f)df}$. Since this integral can generally not be performed analytically we approximate it numerically (for each promoter and probe) by a sum over 100 equal-sized bins of size 0.01 (given the relatively small number of samples per promoter this bin-size is always small compared to the width of the distribution over *f*).

As shown in **Supplementary Figure 7a** the FOVs we observe for CAGE promoters are clearly lower than the FOVs observed for Illumina probes. That is, the expression profiles of CAGE promoters typically vary more across replicates than the expression profiles of micro-array probes. One contributing factor is the limited depth of the CAGE sequencing. That is, CAGE measures a much larger number of independent expression profiles than the micro-array, and many of the CAGE promoters have low overall expression. Because of the Poisson sampling noise in CAGE sequencing, promoters with low expression will generally show noisier expression profiles. Since deepCAGE is a relatively new technology, we currently have only limited insight into other factors that may contribute to noise in the expression profiles. One possible contributing factor is the addition of barcodes to the CAGE tags, as we have observed that replicate samples using different barcodes show larger variations than replicates using the same barcodes (data now shown).

2.5 Construction of position specific weight matrices

For a number of reasons regarding data quality and annotation ambiguities, the construction of a set of position-specific weight matrices (WMs) for human transcription factors is rife with problems that, in our opinion, do not currently have a clean solution. Therefore, our procedures necessarily involve several subjective choices, judgments, and hand-curation, which are certainly far from satisfactory. Our main objectives were:

1. To remove obvious redundancy, we aim to have no more than 1 WM representing any given TF, and where multiple TFs have WMs that are indistinguishable or when their DNA binding domains are virtually identical, then we use only one WM for that set of TFs.
2. Associate WMs with TFs based on the sequences of their DNA binding domains. That is, we obtain lists of TFs that can plausibly bind to the sites of a given WM by comparison of DNA binding domain sequences of TFs known to bind to the sites with those of all other TFs.
3. Re-estimation of WMs using genome-wide predictions of regulatory sites in the proximal promoters of CAGE TSSs.

The input data for our WM construction consists of

1. The collection of JASPAR vertebrate WMs plus, for each WM, the amino acid sequence of the TF that JASPAR associates with the WM.

2. The collection of TRANSFAC vertebrate WMs (version 9.4)

3. The amino acid sequences of all vertebrate TFs in TRANSFAC that are associated with those WMs.

4. A list of 1322 human TFs (Entrez gene IDs) and their amino acid sequences (from RefSeq).

5. A list of 483 Pfam IDs corresponding to DNA binding domains and their Pfam profiles[18].

We start by removing the most basic redundancy from TRANSFAC. TRANSFAC often associates multiple WMs with a single human TF. Although there undoubtedly are cases where a single TF can have multiple distinct modes of binding DNA, and should therefore be realistically represented by multiple WMs, we believe that for the very large majority of TFs it is more realistic to describe the DNA binding specificity of the TF with a single WM. Indeed, a manual inspection of cases in which TRANSFAC associated multiple WMs with a single TF shows that these WMs are typically highly similar and appear redundant. Therefore, for each TF with multiple WMs in TRANSFAC we choose only a single `best' WM based on TRANSFAC's own matrix quality annotation, or WM information score when there were multiple WMs with the same quality score.

Next we ran Hmmer with the DNA binding domain (DBD) profiles from Pfam to extract the DBDs from all transcription factors (E-value cut-off 10^{-9}) associated with either JASPAR or TRANSFAC matrices. We then replaced each such TF with the union of its DNA binding domain sequences. Next we used BLAT to map the DBDs of all TFs associated with JASPAR or TRANSFAC matrices against the entire protein sequences of all human TFs. For each human TF we then extracted a list of all JASPAR/TRANSFAC matrices for which the DBDs of at least one associated TF has a significant BLAT hit (default parameters) against the TF sequence. For each human TF the associated WMs were ordered by the percent identity of the hit, i.e. the fraction of all amino acids in the DBDs that map to matching amino acids in the TF. From this we create a list of `necessary WMs'. For each human TF we obtain the JASPAR WM with the highest percent identity. If there is a TRANSFAC WM with a higher percent identity than any JASPAR TF we record this WM as well. Thus, the necessary WMs are those that are the best match for at least one human TF. This list yielded 381 WMs representing 980 human TFs (often the same WM is the best match for multiple TFs). Manual inspection indicated that a lot of redundancy (essentially identical looking WMs) remained in this list. First we often have both a TRANSFAC and a JASPAR WM for the same TF and moreover often there are multiple TFs, each with its own WM, that look essentially identical. We thus want to fuse WMs in the following situations

1. Different WMs for TFs with identical or near identical DBDs.
2. WMs that are statistically indistinguishable, predict highly overlapping sets of sites, and are associated with TFs that have similar DBDs.

For each pair of WMs we obtain three similarity measurements

1. The percent identity of the DBDs of the TFs associated with the WMs. If there are multiple TFs associated with a WM we take the maximum over all TF pairs.
2. The overlap of the binding sites predicted by each WM. We use MotEvo as described in the methods to predict TFBSs in all proximal promoters and we calculate what fraction of predicted TFBS positions are shared between the sites predicted by the two WMs.
3. A statistical measure of the similarity of the two WMs. Here we take the two sets of sites that define the WMs and calculate the likelihood-ratio of the sets of sites assuming they derive from a single underlying WM and assuming the set of sites for each WM derives from an independent WM.

For each of these three criteria we set a cut-off: 95% identity of the DBDs, 60% overlap of predicted TFBSs, and a likelihood-ratio of exp(40). Using single-linkage clustering, we cluster all WMs whose similarity is over the cut-off for at least 1 of these three criteria. The resulting clusters were then all checked manually and whenever the linkage was dubious we split the cluster. That is, we took a conservative attitude towards removing redundancy and only kept clusters when we were convinced the WMs were essentially identical. For each cluster we then constructed a new WM by aligning the WMs in the cluster and calculating the sum of the base-counts in each column. For a few TFs we obtained more recent WMs from the literature (SP1, OCT4, NANOG, SOX2) and we used these to replace the corresponding WM in the list. For PU.1 we inferred a new WM from the top 50 target regions according to our ChIP-chip data.

Finally, we used MotEvo to predict TFBSs for all WMs in the multiple-species alignments of all human proximal promoters. We then constructed *new* WMs from the list of predicted TFBSs for each WM, weighing each predicted site with its posterior probability (which incorporates the position-specific prior probabilities) and using only sites with a posterior probability of at least 0.5. Our final list contains 201 WMs. For each final WM there is an ordered list of associated human TFs, ordered by percent identity of the DBDs of TFs known to bind sites of the WM and the DBDs of the human

TF. We then checked this list of associations by hand and for each WM cut-off the list of associated human TFs manually. In total 342 human TFs are associated with our 201 WMs. The entire set of WMs and mapping to associated TFs is available from the SwissRegulon website (http://www.swissregulon.unibas.ch).

2.6 Permutation and Cross-validation tests

We tested the significance of the fits using the following permutation test: We randomly permuted the association between the site-counts N_{pm} and the expression profiles e_{ps} so that each promoter is now assigned the site-counts from a randomly chosen other promoter. The model was then fitted on this randomized data set and the fraction of expression signal explained by the fit was calculated exactly as for the original data. This procedure was repeated 1,000 times. For the CAGE promoters, the average fraction of expression signal explained was 0.015 with a standard-deviation of 0.00054, corresponding to a difference of 84.5 standard-deviations with the fitted fraction on the real data (0.061). Assuming the fitted fractions for the permuted data-sets are Gaussian distributed this would correspond to a p-value of 2.85 * 10^{-1554}.

For the cross-validation test we randomly divided the promoters in 10 subsets of equal size. For each subset we use the remaining 90% of the promoters to fit motif activities and used these to predict the expression values of the promoters in the set. Combining the results from all 10 subsets we again calculated the fraction of expression signal explained by the fit. Cross-validation was also applied to the data-set with permuted promoters.

For comparison of the fits based on CAGE versus RefSeq promoters we selected all microarray probes that intersect a RefSeq transcript and that are one-to-one associated with a CAGE promoter region. We fitted the expression data of all these probes once using the site-counts N_{pm} from the associated CAGE promoters and once using N_{pm} from the RefSeq promoters.

2.7 Combing motif activities from replicates and motif FOV

For each motif *m* and each sample *s* our inference provides a fitted activity A^*_{ms} and its associated standard-error σ_{ms} . Therefore, if we ignore covariances between the inferred activities of different motifs, the posterior distribution for the activity of motif *m* in sample *s* is given by $P(A_{ms}) = \frac{1}{\sqrt{2\pi}\sigma_{ms}} \exp\left(-\frac{1}{2}\left(\frac{A_{ms} - A^*_{ms}}{\sigma_{ms}}\right)^2\right)$. For each of the 6 time points we have 6 independent posterior distributions of motif activity, namely 3 replicates for both CAGE and microarray data. We now infer an overall motif activity by combining the 6 posterior distributions. Let's focus on a single motif and let α_t denote the final inferred activity of the motif at time *t*, let C^i_t be the inferred activity from CAGE replicate *i*, σ^i_t its standard-error, M^i_t the inferred activity from microarray replicate *i*, and τ^i_t its standard-error. The posterior distribution for α_t is now given by

$$P(\alpha_t \mid C, M, \sigma, \tau) \propto \prod_{\mathrm{i}} \exp\left(-\frac{1}{2}\left(\frac{\alpha_t - C^i_t}{\sigma^i_t}\right)^2 - \frac{1}{2}\left(\frac{\alpha_t - M^i_t}{\tau^i_t}\right)^2\right) \propto \exp\left(-\frac{1}{2}\left(\frac{\alpha_t - \alpha^*_t}{\sigma^*_t}\right)^2\right),$$ with

$$\alpha^*_t = \frac{\sum_{i=1}^{r} C^i_t(\sigma^i_t)^{-2} + M^i_t(\tau^i_t)^{-2}}{\sum_{i=1}^{r}(\sigma^i_t)^{-2} + (\tau^i_t)^{-2}}$$ and $\sigma^*_t = \left[\sum_{i=1}^{r}(\sigma^i_t)^{-2} + (\tau^i_t)^{-2}\right]^{-1/2}$. That is, the posterior distribution is again Gaussian but with updated mean and standard-error. Finally, we calculate a z-value for the combined activity profile of the motif $z_m = \sqrt{\frac{1}{6}\sum_{t=1}^{6}\left(\frac{\alpha^*_t}{\sigma^*_t}\right)^2}$.

For each motif we also quantify the extent to which the different replicates and measurement technologies (CAGE and microarray) lead to the same inferred activity profiles. For this we calculate a FOV exactly as described in the previous section.

2.8 Clustering motifs on activity profiles

We noticed the inferred activity profiles of several motifs are highly similar suggesting there are clusters of motifs with essentially the same activity profiles. We thus devised a

clustering procedure that joins together motifs whose inferred activity profiles are statistically indistinguishable. To this end we need to calculate, for any set *C* of motifs, the probability of the data under the assumption that their inferred activity profiles all derive from a common underlying activity profile. Let α^{*}_{mt} denote the inferred combined activity of motif *m* at time *t*, let σ^{*}_{mt} denote the standard-error associated with this activity, let *C* denote a cluster of motifs, and let γ_t be the (unknown) common activity profile of the motifs in the cluster. The probability of the inferred activities given γ and the standard-errors is then given by $P(\alpha \mid \gamma, \sigma) = \prod_{m \in C} \left[\prod_t \frac{1}{\sqrt{2\pi}\sigma^{*}_{mt}} \exp\left(-\frac{1}{2} \left(\frac{\alpha^{*}_{mt} - \gamma_t}{\sigma^{*}_{mt}} \right)^2 \right) \right]$. We now use a prior over the underlying activity profile γ that is the same as we used for inferring the activity profiles from the independent data-sets, i.e. $P(\gamma_t \mid \tau) = \frac{1}{\sqrt{2\pi\tau}} \exp\left(-\frac{1}{2} \left(\frac{\gamma_t}{\tau} \right)^2 \right)$, where we again use $\tau = 0.1$. By integrating over the unknown activity profile γ we then obtain the probability of the inferred activity profiles in the cluster under the assumption that they are all the same up to noise, i.e
$P(\alpha \mid \sigma) = \prod_{m \in C} \left\lfloor \int \prod_t P(\alpha^{*}_t \mid \gamma_t, \sigma^{*}_t) P(\gamma_t \mid \tau) d\gamma_t \right\rfloor$. These integrals are all Gaussian integrals and can be performed analytically.

We use the result to hierarchically cluster the 30 core motifs based on their activity profiles. We start with letting each motif be a cluster by itself and calculate, for each pair, the likelihood-ratio of the probability of the data before and after clustering. We then iteratively cluster the pair of motifs with highest likelihood-ratio. Note that when two motifs are clustered we recalculate their average activity profile and associated standard-error of the average exactly in the same way as we do when we combine the data from the replicates (i.e. we treat the inferred activities of the different motifs in the cluster just like we treat the inferred activities from different replicates for the same motif). At each iteration we also keep track of the total probability of the data in the current clustering state. The cut-off for termination of the hierarchical clustering was chosen by hand (essentially where the first large drop in likelihood of the clustering state is observed).

3.0 ChIP on chip analysis of acetylated lysine 9 of histone H3 (H3K9Ac), PU.1 (SPI1), SP1 and RNA Polymerase II (Initiation Complex)

3.1 Chromatin immunoprecipitation

THP-1 cells were cross-linked with 1% formaldehyde for 10 min, and 125mM glycine in PBS was added. Cross-linked cells were collected by centrifugation and washed twice in cold 1 x PBS. The cells were sonicated for 5~7 min with a Branson 450 Sonicator to shear the chromatin.

Complexes containing DNA bound to histone H3 acetylated at lysine 9 (H3K9Ac) were immunoprecipitated with an antibody against H3K9Ac (07-352, Upstate) by overnight rotation at 4°C. The immunoprecipitated sample was incubated with magnetic beads/Protein G (Dynal) for 1 hr at 4°C followed by one wash with each of (1) Low salt wash buffer (0.1% SDS, 1% Triton X-100, 2mM EDTA, 20mM Tris.HCl (pH8.1), 150mM NaCl), (2) High salt wash buffer (0.1% SDS, 1% Triton X-100, 2mM EDTA, 20mM Tris.HCl (pH 8.1), 500mM NaCl) and (3) LiCl wash buffer (10mM Tris.HCl (pH8.1), 0.25M LiCl, 0.5% NP-40, 0.5% Sodium deoxycholate, 1mM EDTA, and two washes with TE buffer). The antibody-H3K9Ac-DNA complexes were eluted from the magnetic beads by addition of 1% SDS and 100 mM $NaHCO_3$. Beads were vortexed for 60 min at RT. The supernatants were incubated for 3.5 hr at 65°C to reverse the cross-links, and incubated for further 30 min at 65°C in the presence of 20mg/ml RNaseA. To purify the DNA, proteinase K solution was added at a final concentration of 100mg/ml, and the samples were incubated overnight at 45°C, followed by a phenol:chloroform:isoamyl alcohol extraction and ethanol precipitation to recover the DNA.

PU.1 (SPI1), SP1 and RNA Polymerase II (PolII) DNA complexes were likewise immunoprecipitated using antibodies T-21 (Santa-cruz), 07-645 (Upstate), and 8WG16 (Abcam), for PU.1 (SPI1), SP1 and PolII, respectively.

3.2 LM-PCR and measurement of H3K9Ac, PU.1 (SPI1) and SP1 by array hybridization

Immunoprecipitated DNA was blunted using 0.25U/µl T4 DNA polymerase (Nippon

Gene). Linker oligonucleotides (5'-accgcgcgtaatacgactcactataggg-3' and Phosphate-5'-ccctatagtgagtcgtattaca-3') were annealed to the DNA while the temperature was decreased gradually from 99°C to 15°C over 90 min. The blunted immunoprecipitated DNA sample was ligated to the annealed oligonucleotides with 500U of T4 DNA ligase (Nippon Gene). The cassette DNA fragments (45ug/reaction) were amplified with Blend *Taq* Plus (Toyobo) using the linker-specific oligonucleotide 5'-accgcgcgtaatacgactcactataggg-3'. PCR cycling conditions were as follows: denaturation at 95°C for 1 min; 25 cycles of 95°C for 30 s, 55°C for 30 s, 72°C for 2 min; and a final extension at 72°C for 7 min. Amplified DNA was purified, fragmented with DNase I (Epicentre), and end-labeled with biotin-ddATP using terminal deoxytransferase (Roche). Amplified DNA was hybridized to Affymetrix whole genome tiling or promoter arrays for 18 h at 45°C, washed, and scanned using the Affymetrix GeneChip System. Each sample was hybridized in triplicate. Affymetrix Human Tiling Arrays (1.0) were used to measure H3K9Ac enrichment. PU.1 (SPI1) and SP1 enrichment were measured using Affymetrix Human Promoter arrays (1.0R). Three technical replicates were performed for ChIP-chip experiments of H3K9, SP1 and SPI1, and two technical replicates for those of PolII.

3.3 IVT and measurement of PolII enrichment by array hybridization

RNA Polymerase II-immunoprecipitated DNA was treated with CIP and poly-dT tailed using terminal transferase. The T7 poly-A primer (5'-CATTAGCGGCCGCGAAATT AATACGACTCACTATAGGGAGAAAAAAAAAAAAAAAAAA [C or T or G] -3') was annealed and the DNA sample was subjected to second strand synthesis using DNA polymerase I (Invitrogen) as follows; 94°C for 2min, ramp down to 35°C (1°C/sec), hold at 35°C for 2 min, ramp down to 25°C (0.5°C/sec), hold and add DNA polymerase I at 37°C for 90 min. After second strand synthesis, the reaction was terminated by EDTA addition and the DNA was column-purified. DNA was amplified by *in vitro* transcription (IVT) using CUGA T7-RNA polymerase (Nippon gene). RNA obtained from poly-dT-tailed DNA was purified using the RNeasy Mini kit (Qiagen) and used to synthesize (cDNA) with SuperScriptII (Invitrogen) and random primers. The DNA T7-polyA primer was annealed to the first strand DNA to synthesize second strand DNA. The second strand DNA was amplified in a second round of IVT, performed as described above. The amplified RNA (cRNA) was also purified in the IVT amplification. The collected cRNA was used to synthesize double-strand cDNA. The double-stranded cDNA, fragmented with DNase I (Epicentre), was end-labelled with biotin-ddATP by using terminal deoxytransferase (Roche). After hybridizing the end-labelled DNA fragments to the tiling arrays (Affymetrix Human Tiling Array 2.0R) for 18 h at 45°C, the arrays were washed and scanned using the Affymetrix GeneChip System. Each of the treatment and control samples was hybridized twice, to provide technical replicates.

3.4 Analysis of Affymetrix tiling array data

The enrichment of DNA fragments immunoprecipitated with H3K9Ac compared to the human genome was determined using the Affymetrix whole-genome tiling array (1.0R). This array tiles the non-repetitive portion of the human genome at 35-bp intervals with more than 41 M pairs of 25-mer probe sequences. The hybridization intensities (background-subtracted intensity; PM – MM, where PM and MM indicate intensities detected by a 25-mer perfectly matching and another one-base-mismatching the genome, respectively) of the probes were measured in three technical replicates and quantile-normalized for each of the treatment and control samples. A shift of the intensities in the treatment relative to control data in a 400-bp window centered at each probe was evaluated by a Wilcoxon Rank Sum test, which assigned a *P*-value to the probe position. We used the Affymetrix software, GTAS (http://www.affymetrix.com/support/developer/downloads/TilingArrayTools) for the *P*-value calculation. Enrichment of DNA fragments precipitated with RNA PolII compared to the human genome was measured by using Affymetrix Human tiling array (2.0R). This array tiles the same portion of human genome as 1.0R with only PM probes. Two technical replicates were performed for both treatment and control samples in measurement of the PolII enrichment, and the enrichment measure, *P*-value was calculated by using GTAS as described for H3K9Ac.

Enrichment of PU.1 (SPI1) and SP1-precipitated DNA was measured using the Affymetrix Human Promoter arrays that tile promoter regions (7.5 kb upstream and 2.45 kb downstream of transcription start sites) of annotated genes at 35-bp intervals with 25-mer probes. Hybridization intensities were measured in three technical replicates for each of the treatment and control samples. The enrichment measure expressed as a *P*-value was calculated by using GTAS as described above.

The genome coordinates of the 25-mer probes, originally based on the version hg16 of human genome, were converted to hg18. The positions of the probes on hg18 were determined by aligning the probe sequences to the human genome (hg18) using Vmatch (http://www.vmatch.de).

SUPPLEMENTARY NOTES

1. ChIP-chip analysis of H3K9 acetylation and RNA Polymerase II and PU.1 (SPI1) binding

1.1. Acetylation and RNA Polymerase II

Whole genome tiling array ChIP-chip data for H3K9 acetylation and Pol II binding (described in **Supplementary Methods**) was analyzed with respect to CAGE promoters. For each promoter, the average signal per tiling array probe in a 2 kb window around the promoter was calculated and compared to the background signal (average signal of all probes on the array, see below), for the PMA 0 and 96 h timepoints separately. For the H3K9 acetylation experiment where two biological replicates were available, the average across both replicates was calculated. In order for a promoter to be included in downstream analysis, a minimum of 10 tiling array probes were required to be present in the window for each promoter; 98% of the promoters met this probe number criterion.

Probe signal is here defined as -log10(p_value) (see **Supplementary Methods**), whereas the background is defined as the average array probe signal plus one or two standard deviations, representing weak and strong association, respectively. For the H3K9 arrays, the background was selected from the biological replicate with the highest average and standard deviation.

Using the above criteria, 62% and 54% of CAGE promoters had strong support from H3K9 acetylation and RNA Polymerase II binding, respectively, in one or both timepoints. With weak criteria, an additional 17% (24%) of the promoters were associated with H3K9 acetylation (RNA Polymerase II binding).

Additionally, H3K9/Pol II enrichment in promoter regions defined by CAGE was compared to the enrichment in non-active promoters. Non-active promoters are here defined as RefSeq transcription start sites at least 1 kb away from any CAGE promoter. For this set of promoters, association with H3K9/Pol II enrichment was computed as for the CAGE promoters above. CAGE promoters were found to be significantly enriched for H3K9 acetylation and Pol II binding ($p < 10^{-15}$, Fisher's exact test; **Figs. SN-1 and SN-2**).

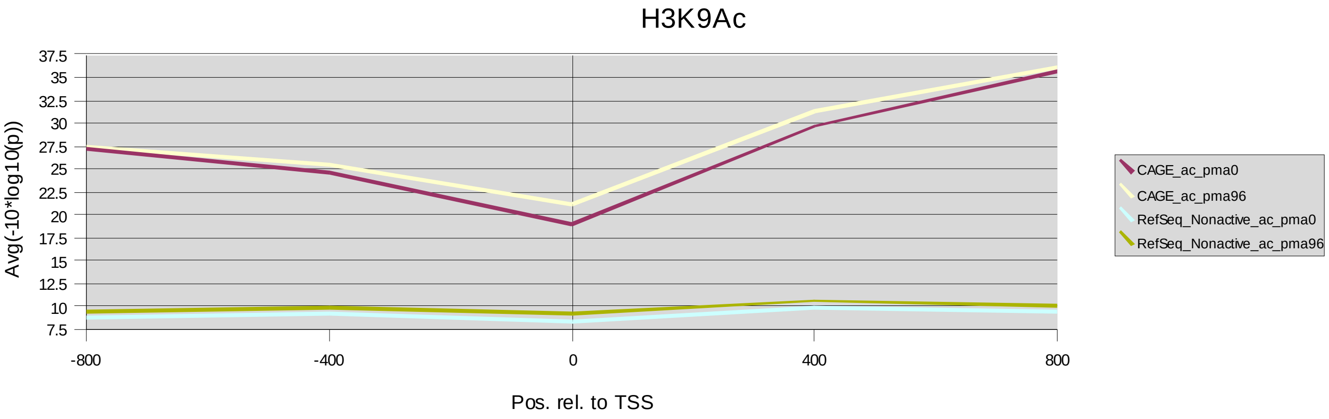


Figure SN-1 Average H3K9 tiling array probe p-value in TSS:s defined by CAGE and non-active RefSeq transcription start sites in each time point.

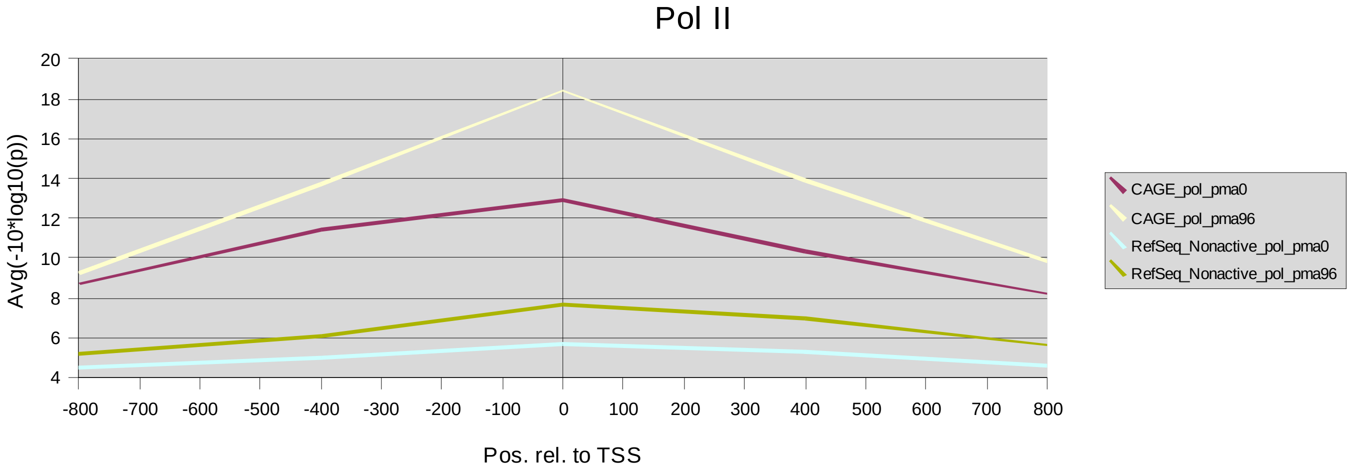


Figure SN-2 Average Pol II tiling array probe p-value in TSS:s defined by CAGE and non-active RefSeq transcription start sites in each time point.

1.2. PU.1 (SPI1)

ChIP-chip for PU.1 binding was performed on Affymetrix Human Promoter arrays and was analyzed with respect to Entrez genes. The entire length of Entrez genes, plus 1 kb upstream, was scanned for PU.1 sites, where a site is defined as a stretch of at least five consecutive tiling array probes having a score (-log10(p_value)) of at least 30. An additional requirement was that no two probes in a site were spaced more than 150 bp apart. This was done separately in each replicate and timepoint.

Using these criteria, 7541 Entrez genes exhibited PU.1 binding in the PMA 0 hour time point, in one or both replicates. 4099 of these were detected in both replicates. In the PMA 96 hour time point, PU.1 showed binding in 8550 (one replicate) or 5002 (both replicates) Entrez genes.

2. Reproducibility of deep CAGE expression measurements between replicates

We first demonstrate the reproducibility of deep CAGE measurements at the promoter level by showing that promoters are conserved between replicates. We define a promoter as present in a given replicate if its expression is non-zero in at least one of the time-points. **Figure SN-3** shows a Venn-diagram of the number of promoters present in the three replicates. This figure shows that on average, a promoter found in a given replicate has a 90% probability of also being present in both other replicates.

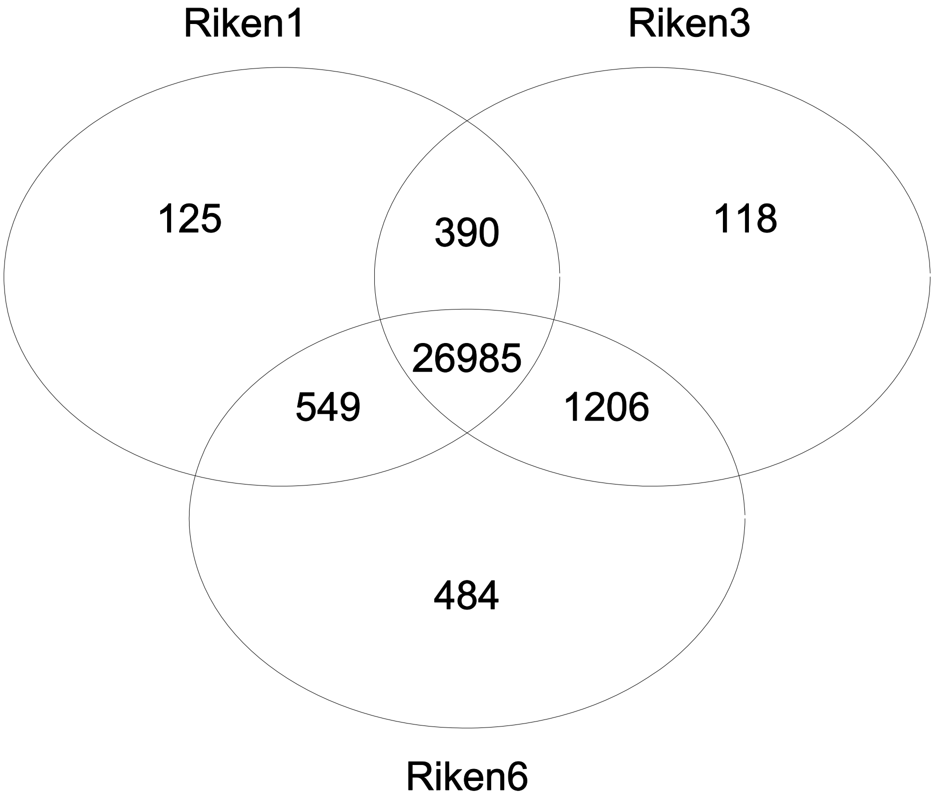


Figure SN-3 The promoters found are consistent between the three replicates.

Next, we compare the expression values of these promoters as measured in the three replicates. **Figure SN-4** shows the scatter plots between the expression values of different promoters in the three replicates for each time point separately. **Table SN-1** shows the corresponding Spearman correlations, which are around 0.55-0.60 in all time points. Between promoter regions, which are less affected by the sampling noise than individual promoters, the Spearman correlations are around 0.65-0.70, as shown in **Table SN-2**.

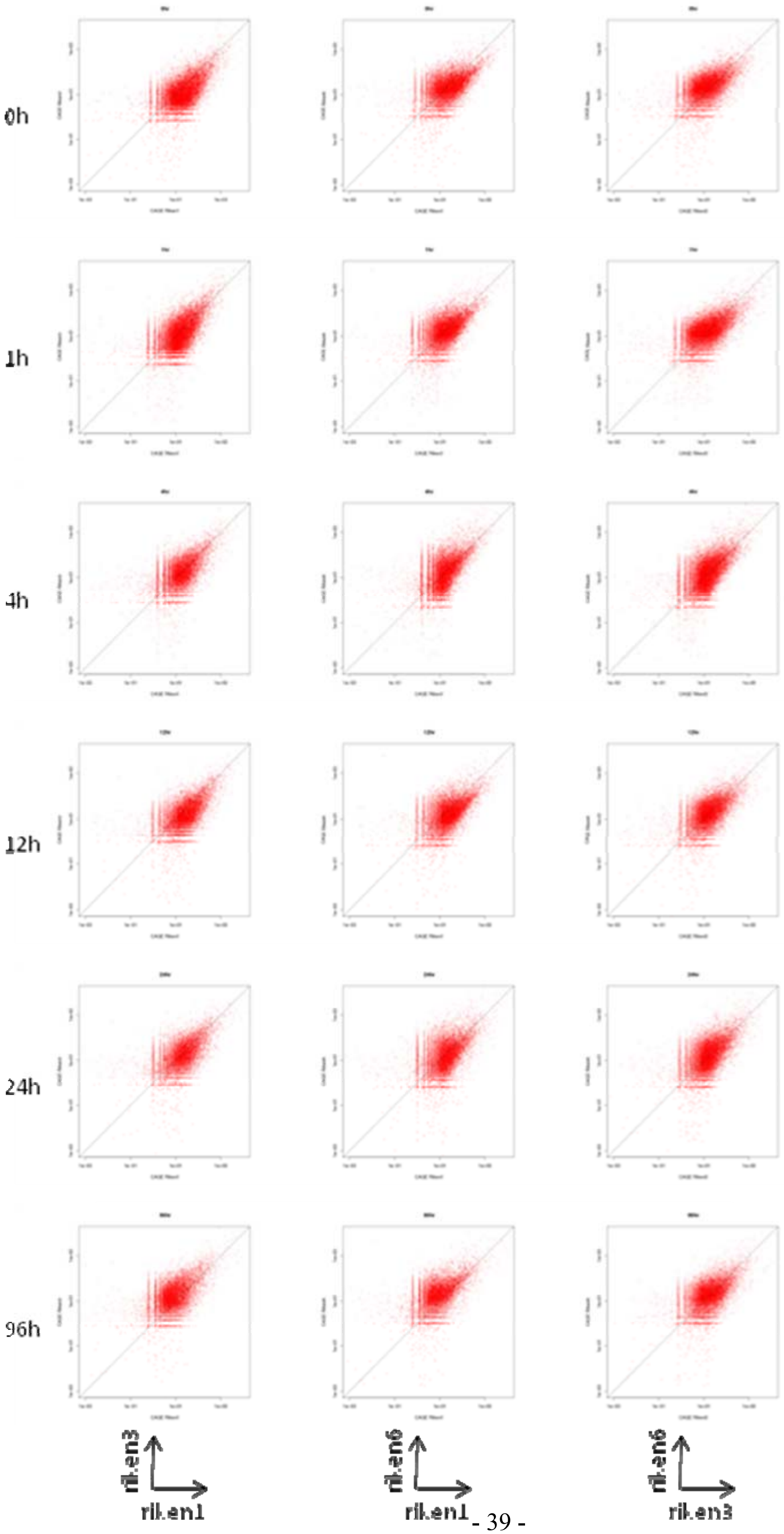

0h
1h
4h
12h
24h
96h
riken3
riken1
riken6
riken1
riken6
riken3

Figure SN-4 Scatter plots of the expression values of promoters measured by the three replicates at the six time points.

Table SN-1 Spearman correlations between the expression values of promoters measured by the three replicates at the six time points.

	RIKEN1-RIKEN3	RIKEN1-RIKEN6	RIKEN3-RIKEN6
0 hr	0.6205	0.5689	0.5893
1 hr	0.6314	0.5964	0.6236
4 hr	0.5378	0.5236	0.6180
12 hr	0.5609	0.5361	0.5746
24 hr	0.5451	0.4751	0.5851
96 hr	0.5424	0.5165	0.5714

Table SN-2 Spearman correlations between the expression values of promoter regions measured by the three replicates at the six time points.

	RIKEN1-RIKEN3	RIKEN1-RIKEN6	RIKEN3-RIKEN6
0 hr	0.7291	0.6767	0.6848
1 hr	0.7462	0.7047	0.7216
4 hr	0.6690	0.6393	0.7122
12 hr	0.6987	0.6528	0.6847
24 hr	0.6641	0.5770	0.6857
96 hr	0.6559	0.6164	0.6683

3. Comparison of deep CAGE to microarray expression profiling

Illumina probes were associated with CAGE promoters and promoter regions as described in the Methods. The diagram in **Figure SN-5** summarizes the association between Illumina probes and CAGE promoter regions. Of the 26608 Illumina probes, 14608 are expressed; 12995 of these intersect a known mRNA and can therefore principle be associated with a CAGE promoter. We find a CAGE promoter region for 9263 of these Illumina probes. For another 2726 Illumina probes, we find CAGE transcription start sites that were not clustered into CAGE-defined promoters. For the remaining 1006 Illumina probes, no associated CAGE expression was found. Likewise, 8297 of the 14607 CAGE–defined promoter regions have an associated Illumina probe.

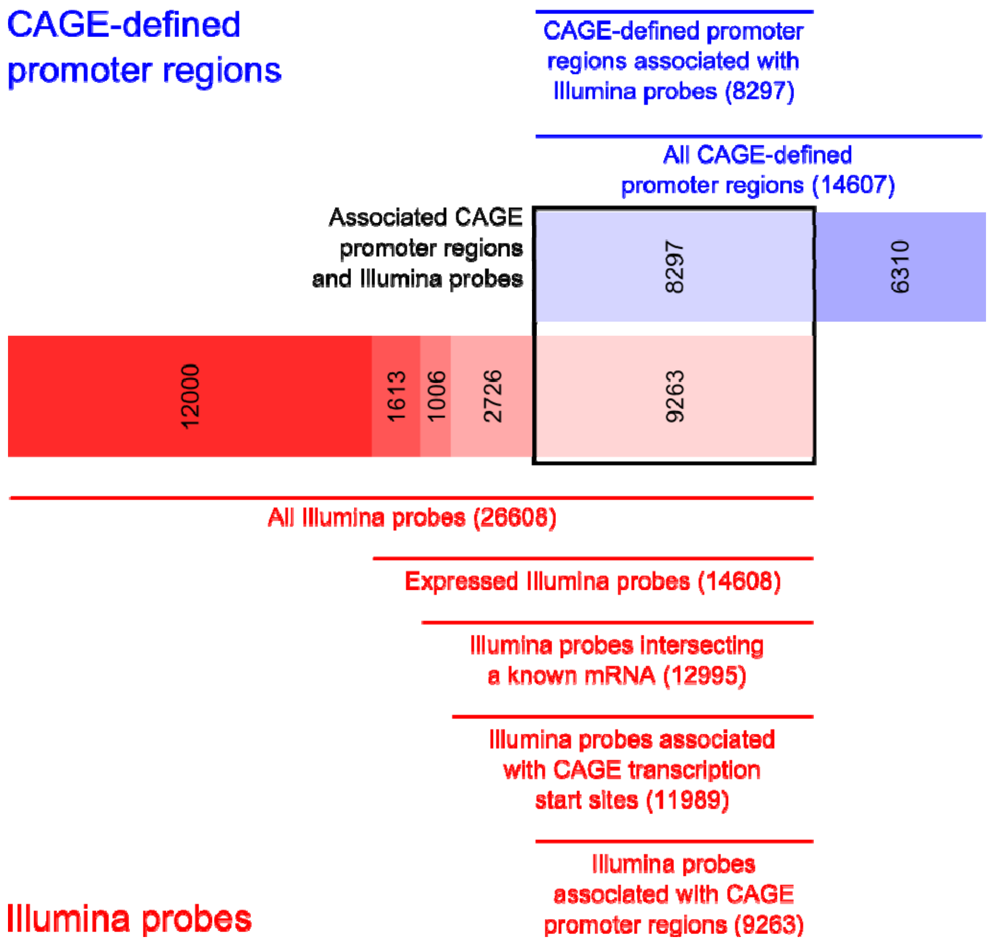


Figure SN-5 Association between Illumina probes and CAGE-defined promoter regions.

Generally, the association between Illumina probes and CAGE promoter regions is not one-to-one. We find multiple CAGE promoter regions associated with 531 Illumina probes, and 1243 CAGE promoter regions are associated with multiple Illumina probes. The comparison of microarray to deep CAGE expression profiling was therefore done on a gene-by-gene basis. As shown in the Venn diagram in **Figure SN-6**, 10690 Entrez genes are expressed in the Illumina microarray experiments, and CAGE expression was found for 9026 Entrez genes; for 7919 Entrez genes we found both CAGE and Illumina expression.

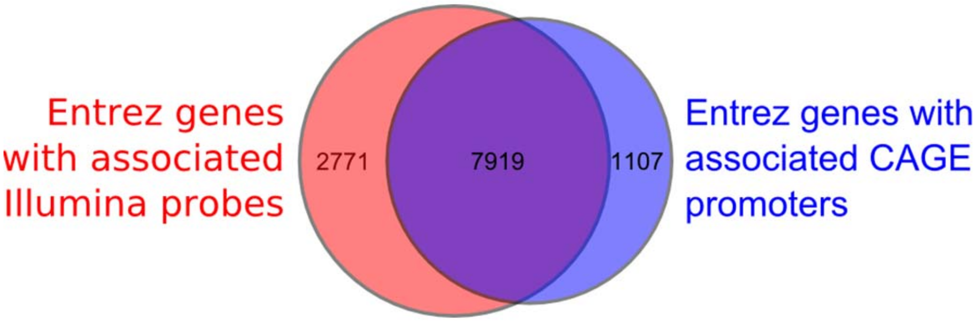

Entrez genes with associated Illumina probes
2771
7919
1107
Entrez genes with associated CAGE promoters

Figure SN-6 Entrez genes expressed in Illumina microarray profiling and deep CAGE expression profiling.

Figure SN-7 shows scatter plots of the CAGE expression log-ratios against microarray expression log-ratios for each of the three replicates and the six time points. The CAGE expression log-ratio was calculated by summing the expression of the CAGE promoter regions associated with a gene, adding 0.5 tpm to each time point in each replicate, taking the logarithm, and subtracting the mean over the time points of each gene in each replicate. The Spearman correlations for these scatter plots are shown in **Table SN-3**; on average, we find a correlation of 0.55.

The value of the Spearman correlations is strongly affected by lowly expressed transcripts, and increases if only genes are included that have a raw CAGE tag count larger than some threshold. **Figure SN-8** shows the Spearman correlation as a function of the threshold, demonstrating that the correlation increases to between 0.7 and 0.9 for all time points and all replicates if we only include genes with at least 100 tags.

Table SN-3 Spearman correlations between the expression values of Entrez genes measured by deepCAGE and Illumina microarray profiling in the three replicates at the six time points.

	RIKEN1	RIKEN3	RIKEN6
0 hr	0.6087	0.5622	0.6524
1 hr	0.6043	0.4949	0.6275
4 hr	0.4954	0.4726	0.6063
12 hr	0.4954	0.2996	0.5669
24 hr	0.5429	0.4801	0.6347
96 hr	0.6126	0.4652	0.7034

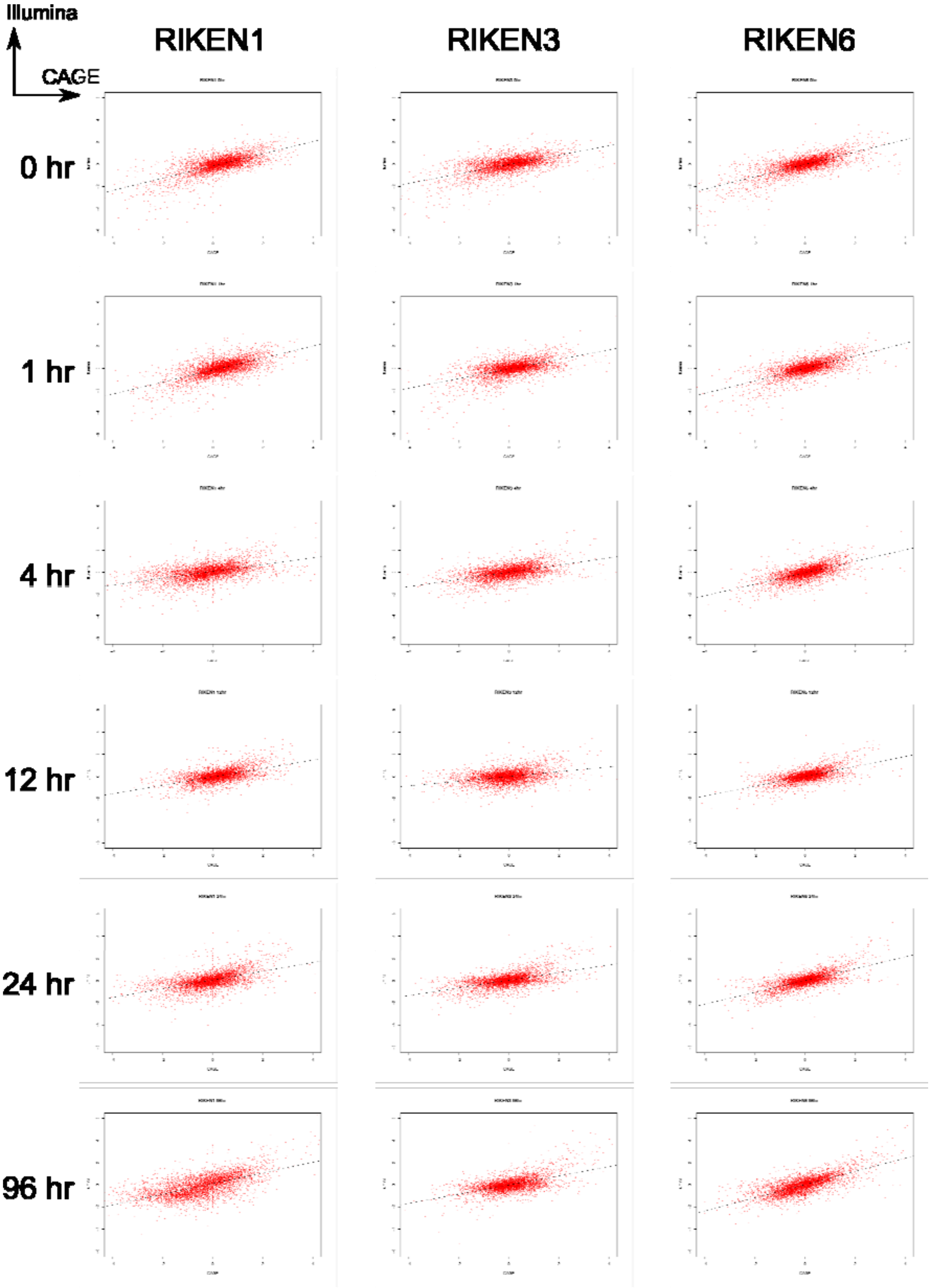

Illumina
CAGE
RIKEN1
RIKEN3
RIKEN6
0 hr
1 hr
4 hr
12 hr
24 hr
96 hr

Figure SN-7 Scatter plots of the expression log-ratios measured by deepCAGE and by Illumina microarray expression profiling in the three replicates for each the time points.

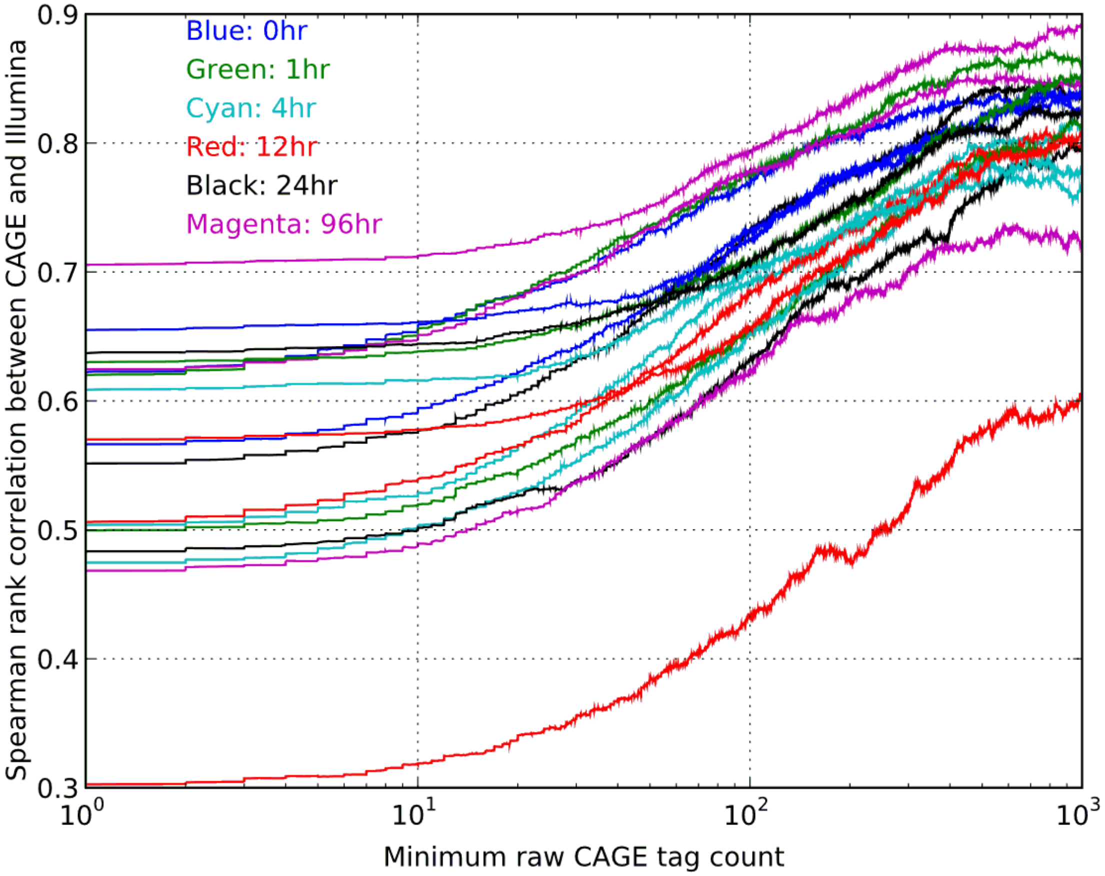
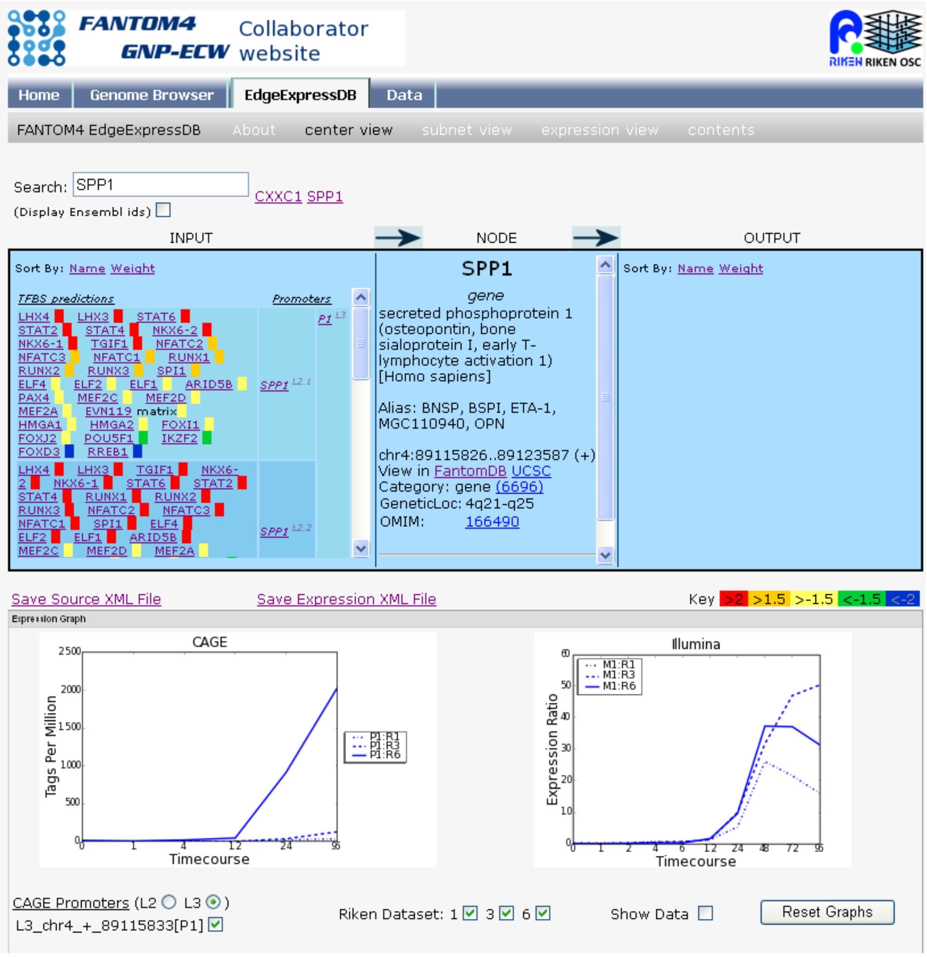


Figure SN-8 Spearman rank correlation between the expression measured by deepCAGE and Illumina expression profiling as a function of the threshold on the raw CAGE tag count. Each curve corresponds to the data of one time point in one replicate. The expression comparisons are performed on a gene-by-gene basis.

4. The predictive power of the data

To illustrate the predictive power of the data sets and predictions generated in this project, we consider the osteopontin (OPN, SPP1, ETA-1) gene as an example. Osteopontin (OPN) is a multifunctional molecule detected in numerous malignant, inflammatory and autoimmune diseases[19]. It is a secreted adhesive molecule, and it is thought to aid in the recruitment of monocytes-macrophages and to regulate cytokine production in macrophages, dendritic cells, and T-cells. **Figure SN-9** shows the data that can be accessed through our web interface (the FANTOM4 GNP-ECW interface and the SwissRegulon interface). A detailed understanding of transcriptional regulation of this gene could provide an explanation for the functional impact of promoter polymorphisms in humans[20].

a



FANTOM4 GNP-ECW Collaborator website

RIKEN RIKEN OSC

Home | Genome Browser | EdgeExpressDB | Data

FANTOM4 EdgeExpressDB About center view subnet view expression view contents

Search: SPP1 CXXC1 SPP1

(Display Ensembl ids) □

INPUT → NODE → OUTPUT

Sort By: Name Weight

TFBS predictions *Promoters*

LHX4 🟥 LHX3 🟥 STAT6 🟥
STAT2 🟥 STAT4 🟥 NKX6-2 🟥
NKX6-1 🟥 TGIF1 🟥 NFATC2 🟧
NFATC3 🟧 NFATC1 🟧 RUNX1 🟧
RUNX2 🟧 RUNX3 🟧 SPI1 🟧
ELF4 🟨 ELF2 🟨 ELF1 🟨 ARID5B 🟨
PAX4 🟨 MEF2C 🟨 MEF2D 🟨
MEF2A 🟨 EVN119 matrix 🟨
HMGA1 🟨 HMGA2 🟨 FOXI1 🟨
FOXJ2 🟨 POU5F1 🟩 IKZF2 🟩
FOXD3 🟦 RREB1 🟦

SPP1 L2.1 *P1* L3

LHX4 🟥 LHX3 🟥 TGIF1 🟥 NKX6-2 🟥 NKX6-1 🟥 STAT6 🟥 STAT2 🟥
STAT4 🟥 RUNX1 🟥 RUNX2 🟥
RUNX3 🟥 NFATC2 🟥 NFATC3 🟥
NFATC1 🟥 SPI1 🟥 ELF4 🟥
ELF2 🟥 ELF1 🟥 ARID5B 🟥
MEF2C 🟨 MEF2D 🟨 MEF2A 🟨

SPP1 L2.2

SPP1

gene

secreted phosphoprotein 1 (osteopontin, bone sialoprotein I, early T-lymphocyte activation 1) [Homo sapiens]

Alias: BNSP, BSPI, ETA-1, MGC110940, OPN

chr4:89115826..89123587 (+)
View in FantomDB UCSC
Category: gene (6696)
GeneticLoc: 4q21-q25
OMIM: 166490

Sort By: Name Weight

Save Source XML File Save Expression XML File Key >2 >1.5 >-1.5 <-1.5 <-2

Expression Graph

CAGE Promoters (L2 ○ L3 ●)
L3_chr4_+_89115833[P1] ☑

Riken Dataset: 1 ☑ 3 ☑ 6 ☑ Show Data □ Reset Graphs

b

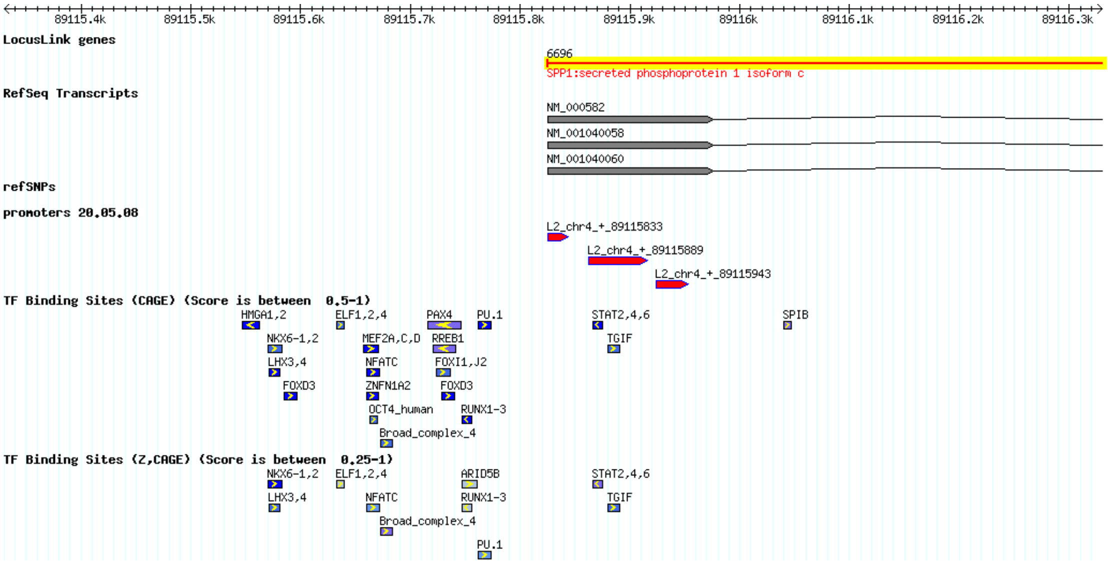


Figure SN-9 (a) A snapshot of the online tool EdgeExpressDB as part of the FANTOM4 web resource and (b) a snapshot of the SwissRegulon web interface.

Figure SN-10 shows the multiple alignment of the promoter region with the orthologous sequences from other mammals onto which the key predicted TFBSs are indicated. The lightly-shaded region is the designated as the 5'UTR, based upon the longest transcript (NM_00582). Although the CAGE tag distribution detects more distal minor sites, it demonstrates clearly that there is a single dominant transcription start site, 30bp downstream of a conserved TATA-like element. This TSS is coincident with the major TSS detected in previous CAGE-based studies of the mouse[21,22]. This finding supports the precise identification of TSS by the deepCAGE methodology.

As shown in **Figure SN-9a**, the CAGE Tag frequency and Illumina microarray data are consistent with each other and show that the *OPN* gene is massively-induced between 12 and 24 hours after addition of PMA, a representative of a cluster of genes that increases in association with development of adherence. We do not predict that this promoter binds MYB within the proximal promoter (**Figs SN-9b and SN-10**). There is a previous report indicating that MYB can activate through a more distal site (at -443), and that binding was confirmed by ChIP in melanoma cells[23]. However, the gene is not expressed in proliferating THP-1 cells, where MYB is highly-expressed. In fact, the gene is induced, albeit weakly, by the MYB knockdown, suggesting that MYB either actually acts as a repressor, or that MYB knockdown activates another TF that activates this gene. Amongst the positive regulators of *OPN* predicted by the network analysis (**Table SN-4** shows the predictions with z-values of 1 or larger), NFATs have not previously been recognised as regulators in THP-1 differentiation, but their role is entirely consistent with their functions in both activated T cells[24] and the fact that OPN was discovered as early T cell activation (ETA-1). The key role of NFATs probably also explains the very high expression of OPN in bone-resorbing osteoclasts, since NFATs are required absolutely for osteoclastogenesis[25].

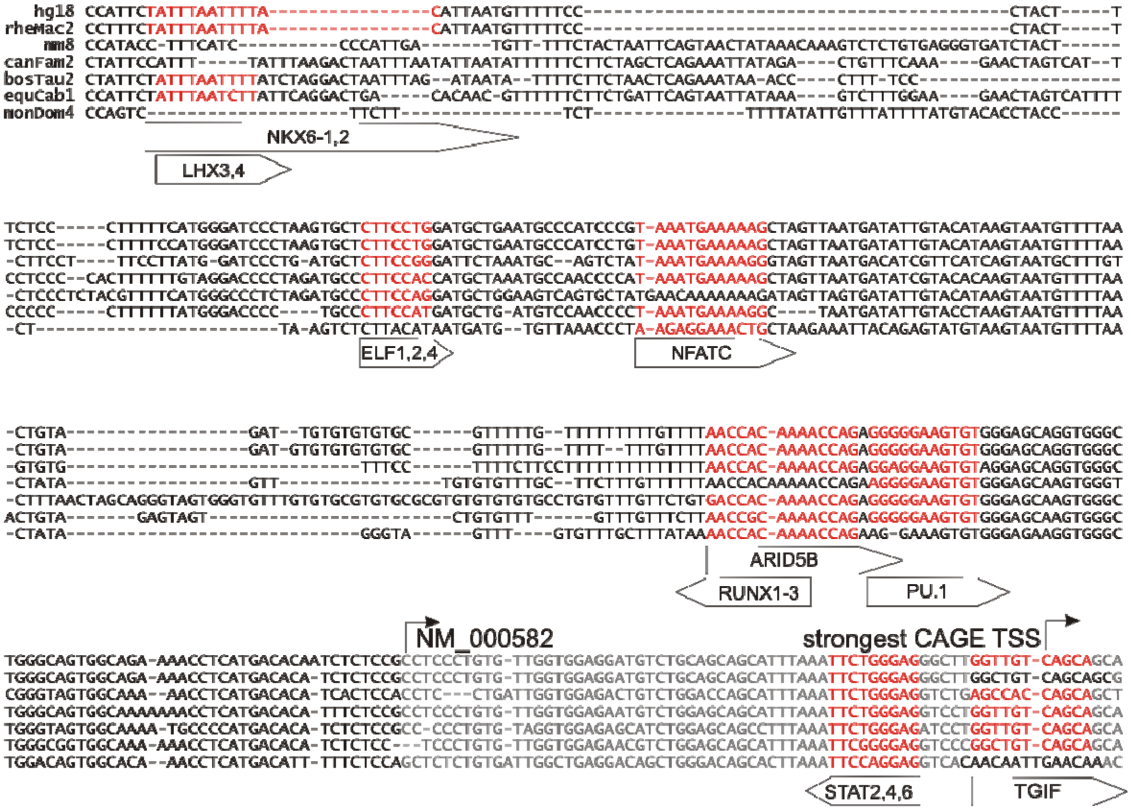


Figure SN-10 Multiple alignment of the *OPN* promoter region.

Table SN-4 Motifs predicted to regulate the main promoter (L2_chr4_+_89115889) of the *OPN* (*SPP1*) gene. The second column shows the z-value for the predicted regulatory interaction.

Motif	z-value
LHX3,4	5.21
TGIF	4.86

NKX6-1,2	4.15
STAT2,4,6	3.75
RUNX1-3	3.44
NFATC	3.17
PU.1	2.99
ELF1,2,4	2.15
ARID5B	2.05

There is well-conserved consensus PU.1 motif in the *OPN* promoter, around 100bp upstream of the TSS. PU.1 (SPI1) has not previously been shown to act upon the *OPN* gene directly, and the promoter architecture is quite different from macrophage-specific genes such as *CSF-1R*, where there are repeated PU.1 sites around a broad TSS, apparently substituting for the TATA box[26]. Since *OPN* is not expressed in undifferentiated THP-1 cells, which express PU.1, it appears that the PU.1 site is not sufficient to generate *OPN* expression. The PU.1 knockdown, followed by PMA treatment shows a modest down-regulation of *OPN* when compared with PMA treatment without PU.1 knockdown. This result suggests that PU.1 induction/activation contributes only partly to the induction of the gene.

The direct binding of RUNX2 to the *OPN* promoter, and key function of this motif, has been confirmed previously[27], in THP-1 cells, we predict that RUNX1 (AML1) is the dominant family member acting through this site, but RUNX2 is also expressed and regulated.

The potential action of members of the STAT family in the regulation of *OPN* has not previously been recognised, and it raises the possibility of an interesting feedback loop. OPN has been identified as a feedback regulator of inflammation that signals the degradation of STAT1[28]. Amongst the factors that were not previously recognised as candidate regulators, TGIF is a homeobox transcription factor previously implicated in myelogenous leukaemia. Although it is considered as a repressor, there are multiple isoforms made from the complex locus[29]. The LHX and NKX6 factors also have no known roles in myeloid biology, and the former are actually not expressed. It therefore seems likely that there are other factors that bind these AT-rich motifs. Amongst the expressed transcription factors, there are several that share the LIM domain with LHX family and would be candidates.

Taken together, the data predict that *OPN* is a target for the concerted actions of a substantial number of different transcription factors activated during cellular differentiation.

SUPPLEMENTARY FIGURE AND TABLE LEGENDS

Supplementary Figure Legends

Supplementary Figure 1 THP-1 clone 5 differentiation with 96 hour PMA treatment. (**a**) qRT-PCR confirmation of CSF1R and APOE induction, and (**b**) Morphology of PMA differentiated THP-1 clone5.

Supplementary Figure 2 Induction of key monocytic differentiation markers in THP-1 subclone 5. (**a**) Array results with PMA and for selected pro-differentiative siRNAs (Note MYB siRNA significantly induces all of these markers). Y axis shows expression ratio relative to 0h for PMA and Negative control for siRNAs. Measurements are from Illumina human genome Sentrix6 microarrays (v2). Average and standard deviation (error bars) of 3 biological replicates are shown. Note similar profiles were generated by deepCAGE for PMA induced changes (data not shown). (**b**) representative FACs profiles for CD14 staining in negative control, MYB and GFI1 knockdowns.

Supplementary Figure 3 Total expression per TSS, promoter, and promoter region. Shown are the distributions of the total number of tags per million (**a**) and total raw number of tags (**b**) across individual TSS (blue curve), promoters (green curve), and promoter regions (red curve). Both axes are shown on logarithmic scales.

Supplementary Figure 4 Distribution of distances between promoters and the start of the nearest known transcript. Note the vertical axis is shown on a logarithmic scale. The inset shows the fractions of promoters within 1Kb of a known start, those further than 1Kb from a known start but within 1Kb of a gene locus, and those distal to gene loci.

Supplementary Figure 5 Average phastCons[30] conservation scores in a 10Kb region around TSSs within 1Kb of known starts (**b**) and around TSSs more than 1Kb away from any known gene locus (**a**).

Supplementary Figure 6 Active promoters in THP-1 differentiation. Red boxes show the promoter regions detected by deepCAGE for the example genes (**a**) DTNA, (**b**) AGPAT1, (**c**) LST1 and (**d**) GFI1. Note, the third promoter* in GFI1 does not map to a full length transcript however there is EST support (BM149905).

Supplementary Figure 7 Reproducibility of the expression profiles across the three biological replicate time series, and correlation between the expression profiles based on CAGE and microarray measurements. (**a**) Distributions of the "expression signal" of the promoters/probes defined as the fraction of expression variance (FOV) that is reproduced across the three replicates. The whiskers denote 5 and 95 percentiles, the bar the 25 and 75 percentiles and the vertical line denotes the median fraction of variance for CAGE promoters that are associated with 1 microarray probe (red), all CAGE promoters (light red), microarray probes associated with 1 CAGE promoter (green) and all microarray probes (light green). (**b**) Distribution of Pearson correlation coefficients of the expression profiles of microarray probes and associated CAGE promoters. Whiskers denote 5 and 95 percentiles, boxes 25 and 75 percentiles and the vertical line the median correlation coefficient for probes associated with 1 CAGE promoter (light blue), probes associated with multiple CAGE promoters (blue), correlations of the replicate-averages for microarray probes associated with 1 CAGE promoter (light brown) and probes associated with multiple CAGE promoters (brown). (**c**) Representative scatterplot of deepCAGE biological replicates for undifferentiated THP-1 cells. (**d**) Representative scatterplot of median normalized log expression ratios for Illumina and CAGE for undifferentiated THP-1 cells (full versions of all comparisons are provided in the **Supplementary Notes**).

Supplementary Figure 8 Positional distribution relative to TSS of predicted TFBSs for the 15 most significant motifs. The horizontal axis shows the position relative to TSS and the vertical axis shows the fraction of all promoters that have a site for the motif centered precisely at the corresponding position.

Supplementary Figure 9 Inferred motif activities across replicates (CAGE and microarray) for the top 10 most significant motifs. Motifs are ordered by significance from top left to bottom right. Each pair of panels corresponds to the activities inferred from CAGE (left) and microarray data (right). The activities inferred for the three biological replicates are shown in red, green, and blue.

Supplementary Figure 10 Fraction of expression signal explained by the motif activities for different data sets under permutation and 10-fold cross validation tests. Different combinations of expression data and TFBS predictions tested were (**a**) expression variance of 29,857 CAGE promoters modeled using TFBS predictions from CAGE defined promoters, (**b**) expression variance of the 8,416 expressed array probes that are associated with both a RefSeq and a CAGE promoter, using TFBSs from CAGE defined promoters, (**c**) expression variance of the same 8,416 array probes using TFBSs from Refseq defined promoters, and (**d**) expression variance of all 11,995 expressed array probes using CAGE TFBS predictions whenever available, and Refseq TFBS prediction when no CAGE promoter was associated with the transcript. For each we determined the fraction of expression signal (expression variance minus variance in replicate noise) that is explained by the model (dark blue), when the association between promoters and expression profiles is randomly permuted (purple/brown), under 10-fold cross-validation (yellow), and under 10-fold cross-validation of the randomly permuted data (light blue). The model explains 6% of the expression signal of all 29,857 promoters, comparable with statistics obtained in recent work[31] for the comparatively simpler task of explaining expression differences between pairs of samples for a selected set of highly varying genes. Comparison of the amount of expression signal explained by the model compared to a data-set in which the assignment between promoters and expression profiles is randomly permuted (1.5% of expression signal explained) demonstrates the extreme significance of the inferred activity profiles (estimated p-value 2.85 * 10^{-1554}). A 10-fold cross-validation test (on average 3.4% explained versus -1.2% 'explained' for permuted promoters in a 1000 iterations, which corresponds to a difference of 170 standard deviations) further demonstrates the validity of the fitting. The fact that the 10-fold cross-validation of the randomized data resulted in negative values indicates that the residual variance after prediction is larger than the original variance. Comparison of the explained expression signal in (**b**) and (**c**), where we considered the 8,416 expressed microarray probes that are associated with both CAGE and RefSeq promoters, demonstrates that the predicted TFBSs in CAGE promoters provide significantly better fits than the TFBSs in the corresponding RefSeq promoters, i.e. 7.8% versus 6.3% of explained expression signal. Note that, because the set of promoters/probes fitted in (**a**), (**b**,**c**), and (**d**) are different, the fractions of expression signal explained cannot be compared across these different data-sets. Only the values in (**b**) and (**c)** can be directly compared.

Supplementary Figure 11 Quality of the fits as a function of various CAGE promoter statistics. (**a**) Mean fraction of expression variance (FOV) explained by the fits as a function of the absolute expression (average log-tpm) of the promoter. (**b**) Mean fraction of expression variance (FOV) explained by the fits as a function of the reproducibility of the promoter's expression profile, as estimated by the fraction of the variance in the promoter's expression profile that is reproduced across the 3 replicates (FOV). (**c**) Mean fraction of expression variance (FOV) explained by the fits as a function of the variance of the promoter's expression profile. (**d**) Blow up of the right half of panel (**c**). For each statistic all CAGE promoters were divided into 10 bins and for each bin the average FOV and its standard-error (shown as error bars) were determined. Note that all FOVs are as determined from a single fit of activities based on the expression of all promoters, i.e. we do not re-estimate motif activities based on different promoter subsets.

Supplementary Figure 12 Quality of the fits at each time point for all replicates. (**a**): Quality of the fits as measured by FOV (Fraction of Variance in the expression across all promoters explained by the fit) for each time point in each of the CAGE replicates. (**b**): Quality of the fits as measured by FOV (Fraction of Variance in the expression across all probes explained by the fit) for each time point in each of the Illumina micro-array replicates.

Supplementary Figure 13 Confirmation of transcription factor knockdown by western blot. Total protein lysate 48 hrs post transfection.

Supplementary Figure 14 Log expression ratio (fold-change) differences of predicted targets and non-targets for several different siRNAs. **(a)** Difference in average log expression ratio upon siRNA knockdown between predicted targets and non-targets as a function of the z-value cut-off on the target prediction for knockdown of SP1. **(b)** Difference in average log expression ratio upon siRNA knockdown between predicted targets and non-targets as a function of the z-value cut-off on the target prediction for knockdowns of PU.1 (SPI1) using two different siRNAs (PU.1 in pink and PU.1_2 in purple). All lines are linear regression fits to the data.

Supplementary Figure 15 Comparison of the cumulative distributions of measured log expression ratio (fold-change) under siRNA knockdowns for predicted targets (red) and predicted non-targets (green) of 8 different transcription factors. Each panel corresponds to a TF, indicated at the top of the panel. Panels (**a**) and (**b**) show two TFs with highly significant down-regulation of predicted targets. Note that for MYB more than 70% of all predicted targets of MYB are down-regulated. Panel (**c**) shows an example TF (PU.1) whose targets are enriched for probes that show large down-regulation. Panel (**d**) shows an example of a TF, SP1, that seems to act as a repressor, i.e. knockdown leads to small but consistent up-regulation of its predicted targets. Panels (**e**) and (**f**) show two examples (YY1 and NFYA) of TFs where the siRNA validation experiment seems to have been unsuccessful, i.e. predicted targets and non-targets do not show significant differences in their log expression ratio (fold-change) distributions. Panels (**g**) and (**h**) show two closely-related TFs, CEBPA and CEBPB, of which only one shows significant differences between predicted targets and non-targets.

Supplementary Figure 16 MYB knockdown in THP-1 cells elicits an adherence phenotype. Images were taken from the bottom of the dish. Note floating cells in non transfected and negative control siRNA samples. "NC-FITC" shows cells transfected with A FITC labeled negative control siRNA.

Supplementary Figure 17 Early differentiation involves proportionally more TFs than non-TFs. (**a**) Using microarrays we count the number of genes with significant differences in expression levels compared to the undifferentiated state. Note: Early differentiation is enriched for changes in TFs. (**b**) Induction and repression of all genes during PMA differentiation (relative to 0h). (**c**) Induction and repression of TFs[15] during PMA differentiation relative to 0h.

Supplementary Tables

Supplementary Table 1 Distribution of the number of promoters per gene (zero counts not shown). We identified 9452 genes with at least one CAGE-defined promoter. The promoters shown in this table account for 24,327 out of the 29,857 promoters identified in total. 300 promoters are associated with two genes and 8 promoters with three genes. The remaining 5530 promoters were not assigned to any gene.

Supplementary Table 2 Distribution of the number of promoters per promoter region (zero counts not shown).

Supplementary Table 3 ChIP-chip enrichment in predicted targets compared to predicted non-targets**.** Public ChIP on chip data were extracted for SRF[32], E2F4, ELF1, ETS1, GABPA, RUNX1[33], YY1[34], E2F1, E2F4, E2F6[35], and MYC[36]. Due to the diverse sources of ChIP data and the methods and thresholds used, we took all edges reported by the respective papers and converted them into a matrix to Entrez geneID relationship (edge). All edge comparisons were then made on the basis of matrix to Entrez geneID, and significance tested using Fisher's exact test. For the SP1 and PU.1 (SPI1) comparisons we used in house ChIP data from hybridizations on Affymetrix promoter tiling arrays (described in the **Supplementary Methods**).

Supplementary Table 4 Significance and reproducibility of the 30 core motifs. The significance is quantified by the overall z-value of the motif (activity relative to its standard-error) and the reproducibility is quantified by the fraction of variance that is reproduced across replicates and measurement technology, i.e. CAGE and microarray. Statistics for all other motifs are available from the FANTOM4 web resource.

Supplementary Table 5 Validation of the core network. For all predicted edges the literature was searched for evidence of protein-DNA binding. In addition high throughput ChIP-chip data from the literature and PU.1 (SPI1) ChIP from this paper was used to validate predicted edges. Finally siRNA perturbation edges, with a B-stat > 0 were considered as validation of predicted edges.

Supplementary Table 6 Gene Ontology terms enriched among predicted target genes (based on both CAGE and Illumina microarray data) of the core 30 motifs.

Supplementary Table 7 Sequences of siRNAs and qRT-PCR primers. Knockdown rate of triplicate experiments is also shown. PU.1 (SPI1).

Supplementary Table 8 All siRNAs tested and their differentiative effect. Focusing only on the set of genes greater than 2 fold up- or down-regulated in the arrays (and with a B-statistic >2.5) we identified 967 genes up-regulated and 916 genes down-regulated in the differentiated state compared to undifferentiated state of the PMA time course. Genes affected upon knockdown were then compared to these lists. Full set of the data is available from the FANTOM4 web resource.

Supplementary Table 9 Fourteen TFs that have differentiative overlap larger than 50%. Also shown are the differentiative overlap with 4 different negative control samples (NC 0 and NCs2, 3, and 4). Third column shows the p-value for the overlap under a permutation test.

Supplementary Table 10 Confirmation of surface marker changes by flow cytometry. Note significant up-regulation of CD11b (ITGAM) with both MYB and GFI1 siRNAs, and up-regulation of CD54 and CD14 with MYB siRNA.

Supplementary Table 11 Relationship between pro-differentiative changes induced by MYB siRNA and other siRNAs.

Supplementary Table 12 Transcription factors detected in THP-1 cells during differentiation.

Supplementary Table 13 The enriched TF motifs in the promoters of TF co-expression clusters. Only the top 10 motifs are shown. All data sets are available from the FANTOM4 web resource.

Supplementary Table 14 Frequency of leukemia related terms in entrez gene annotations for transcription factors down-regulated, up-regulated and transiently induced/repressed during PMA-induced differentiation. (**a**) number of TFs from each class with terms cancer, leukemia, ‘myeloid leukemia’, and lymphoma. (**b**) percentage of TFs from each class with these terms. (**c**) p-value for the observation (background used is all TFs).

Supplementary Table 15 Accession numbers of the datasets in the public databases, DDBJ (DNA Data Bank of Japan) and CIBEX (Center for Information Biology gene Expression database).

REFERENCES FOR SUPPLEMENTARY INFORMATION

1. Eperon, S., De Groote, D., Werner-Felmayer, G. & Jungi, T.W. Human monocytoid cell lines as indicators of endotoxin: comparison with rabbit pyrogen and Limulus amoebocyte lysate assay. *J Immunol Methods* **207**, 135-45 (1997).
2. Shiraki, T. et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci U S A* **100**, 15776-81 (2003).
3. Kodzius, R. et al. CAGE: cap analysis of gene expression. *Nat Methods* **3**, 211-22 (2006).
4. Faulkner, G.J. et al. A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. *Genomics* (2008).
5. Gershenzon, N.I., Stormo, G.D. & Ioshikhes, I.P. Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites. *Nucleic Acids Res* **33**, 2290-301 (2005).
6. Loh, Y.H. et al. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat Genet* **38**, 431-40 (2006).
7. Siddharthan, R., Siggia, E.D. & van Nimwegen, E. PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol* **1**, e67 (2005).
8. Notredame, C., Higgins, D.G. & Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**, 205-17 (2000).
9. van Nimwegen, E. Finding regulatory elements and regulatory motifs: a general probabilistic framework. *BMC Bioinformatics* **8 Suppl 6**, S4 (2007).
10. Moses, A.M., Chiang, D.Y., Pollard, D.A., Iyer, V.N. & Eisen, M.B. MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol* **5**, R98 (2004).
11. Beissbarth, T. & Speed, T.P. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* **20**, 1464-5 (2004).
12. Smyth, G.K., Yang, Y.H. & Speed, T. Statistical issues in cDNA microarray data analysis. *Methods Mol Biol* **224**, 111-36 (2003).

13. Smyth, G.K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3**, Article3 (2004).

14. Lin, S.M., Du, P., Huber, W. & Kibbe, W.A. Model-based variance-stabilizing transformation for Illumina microarray data. *Nucleic Acids Res* **36**, e11 (2008).

15. Roach, J.C. et al. Transcription factor expression in lipopolysaccharide-activated peripheral-blood-derived mononuclear cells. *Proc Natl Acad Sci U S A* **104**, 16245-50 (2007).

16. Margulies, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-80 (2005).

17. Faulkner, G.J. et al. A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. *Genomics* **91**, 281-8 (2008).

18. Wilson, D., Charoensawan, V., Kummerfeld, S.K. & Teichmann, S.A. DBD--taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res* **36**, D88-92 (2008).

19. Scatena, M., Liaw, L. & Giachelli, C.M. Osteopontin: a multifunctional molecule regulating chronic inflammation and vascular disease. *Arterioscler Thromb Vasc Biol* **27**, 2302-9 (2007).

20. Hummelshoj, T., Ryder, L.P., Madsen, H.O., Odum, N. & Svejgaard, A. A functional polymorphism in the Eta-1 promoter is associated with allele specific binding to the transcription factor Sp1 and elevated gene expression. *Mol Immunol* **43**, 980-6 (2006).

21. Carninci, P. et al. The transcriptional landscape of the mammalian genome. *Science* **309**, 1559-63 (2005).

22. Carninci, P. et al. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38**, 626-35 (2006).

23. Schultz, J. et al. The functional -443T/C osteopontin promoter polymorphism influences osteopontin gene expression in melanoma cells via binding of c-Myb transcription factor. *Mol Carcinog* **48**, 14-23 (2008).

24. Cockerill, P.N. Mechanisms of transcriptional regulation of the human IL-3/GM-CSF locus by inducible tissue-specific promoters and enhancers. *Crit Rev Immunol* **24**, 385-408 (2004).

25. Takayanagi, H. The role of NFAT in osteoclast formation. *Ann N Y Acad Sci* **1116**, 227-37 (2007).

26. Ross, I.L., Yue, X., Ostrowski, M.C. & Hume, D.A. Interaction between PU.1 and another Ets family transcription factor promotes macrophage-specific Basal transcription initiation. *J Biol Chem* **273**, 6662-9 (1998).

27. Inman, C.K. & Shore, P. The osteoblast transcription factor Runx2 is expressed in mammary epithelial cells and mediates osteopontin expression. *J Biol Chem* **278**, 48684-9 (2003).

28. Gao, C., Guo, H., Mi, Z., Grusby, M.J. & Kuo, P.C. Osteopontin induces ubiquitin-dependent degradation of STAT1 in RAW264.7 murine macrophages. *J Immunol* **178**, 1870-81 (2007).

29. Hamid, R., Patterson, J. & Brandt, S.J. Genomic structure, alternative splicing and expression of TG-interacting factor, in human myeloid leukemia blasts and cell lines. *Biochim Biophys Acta* **1779**, 347-55 (2008).

30. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-50 (2005).

31. Das, D., Nahle, Z. & Zhang, M.Q. Adaptively inferring human transcriptional subnetworks. *Mol Syst Biol* **2**, 2006 0029 (2006).

32. Cooper, S.J., Trinklein, N.D., Nguyen, L. & Myers, R.M. Serum response factor binding sites differ in three human cell types. *Genome Res* **17**, 136-44 (2007).

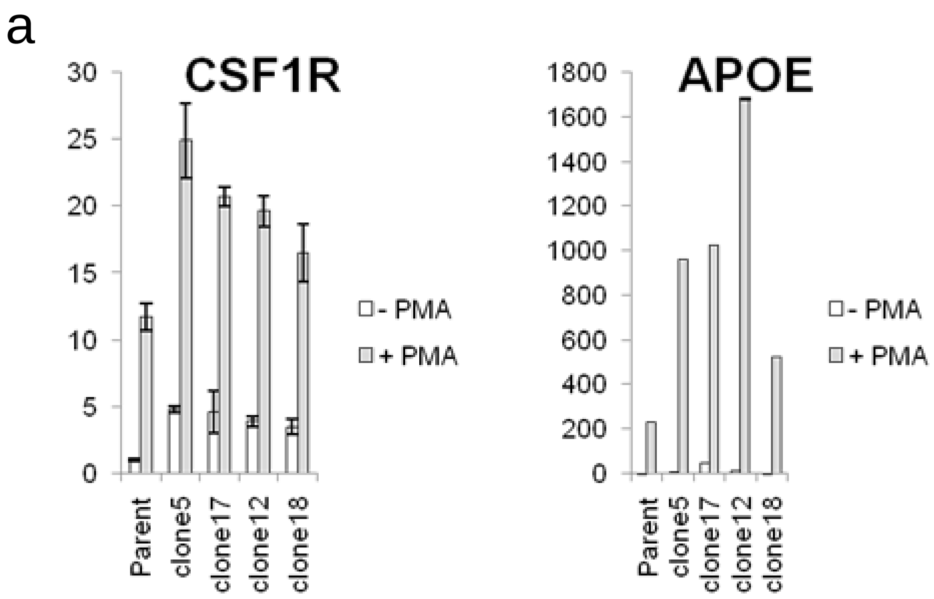
33. Hollenhorst, P.C., Shah, A.A., Hopkins, C. & Graves, B.J. Genome-wide analyses reveal properties of redundant and specific promoter occupancy within the ETS gene family. *Genes Dev* **21**, 1882-94 (2007).

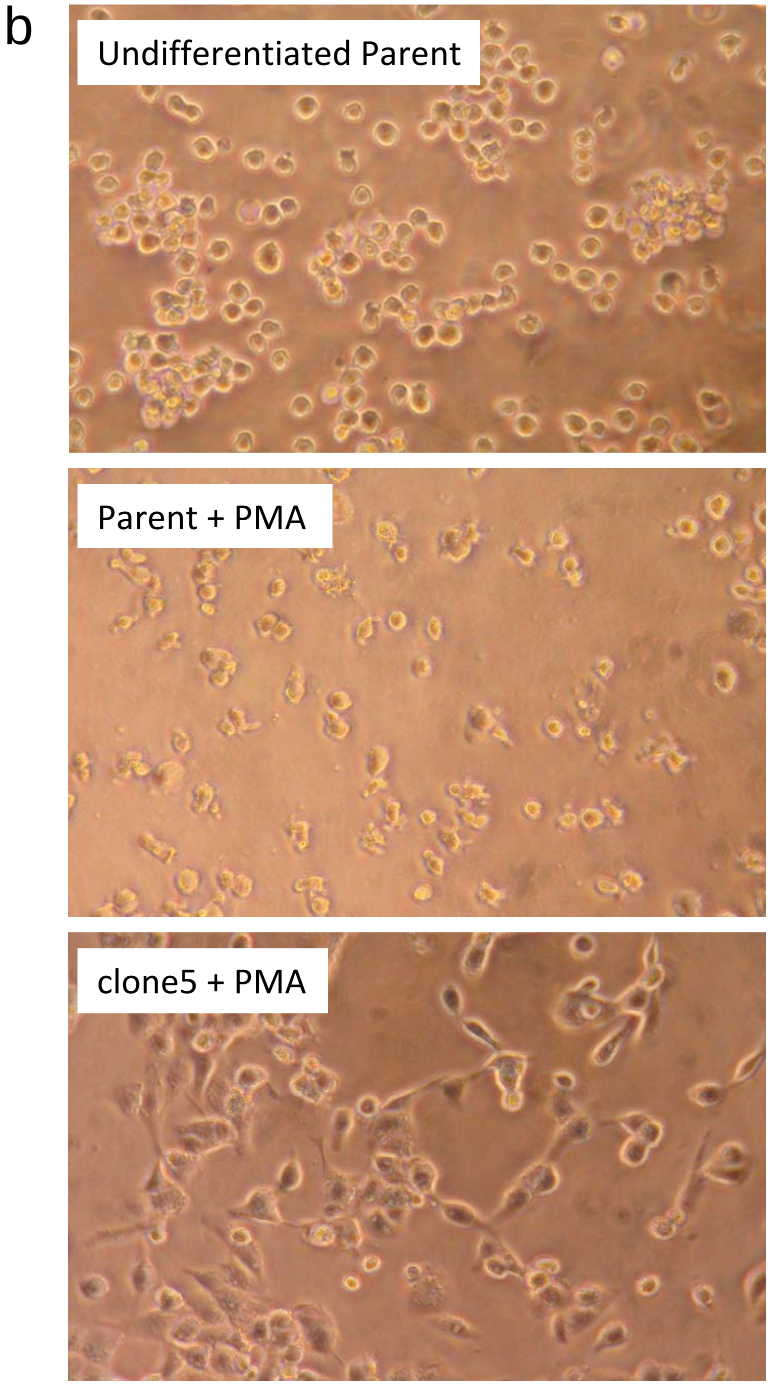
34. Xi, H. et al. Analysis of overrepresented motifs in human core promoters reveals dual regulatory roles of YY1. *Genome Res* **17**, 798-806 (2007).

35. Xu, X. et al. A comprehensive ChIP-chip analysis of E2F1, E2F4, and E2F6 in normal and tumor cells reveals interchangeable roles of E2F family members. *Genome Res* **17**, 1550-61 (2007).

36. Zeller, K.I. et al. Global mapping of c-Myc binding sites and target gene networks in human B cells. *Proc Natl Acad Sci U S A* **103**, 17834-9 (2006).

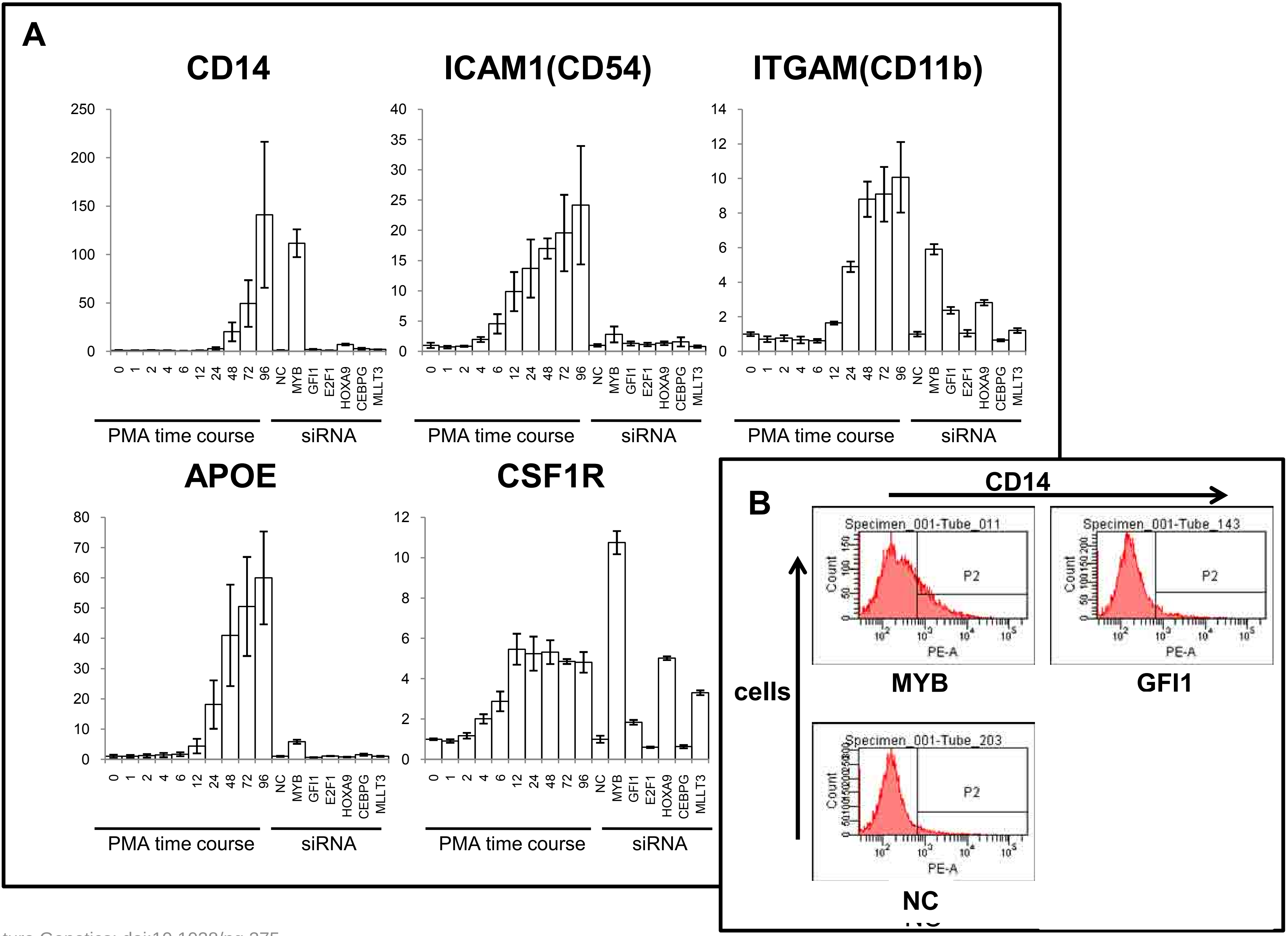
Supplementary Figure 1



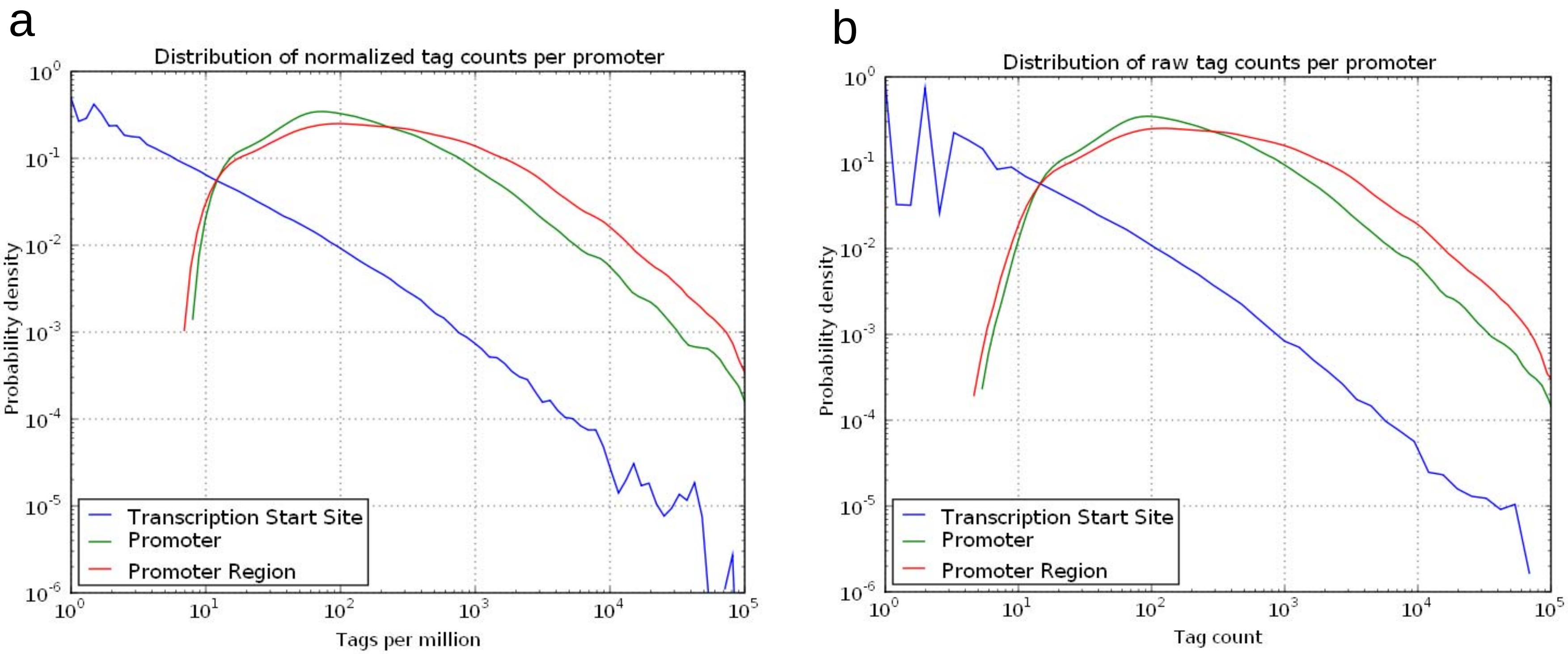


Nature Genetics: doi:10.1038/ng.375

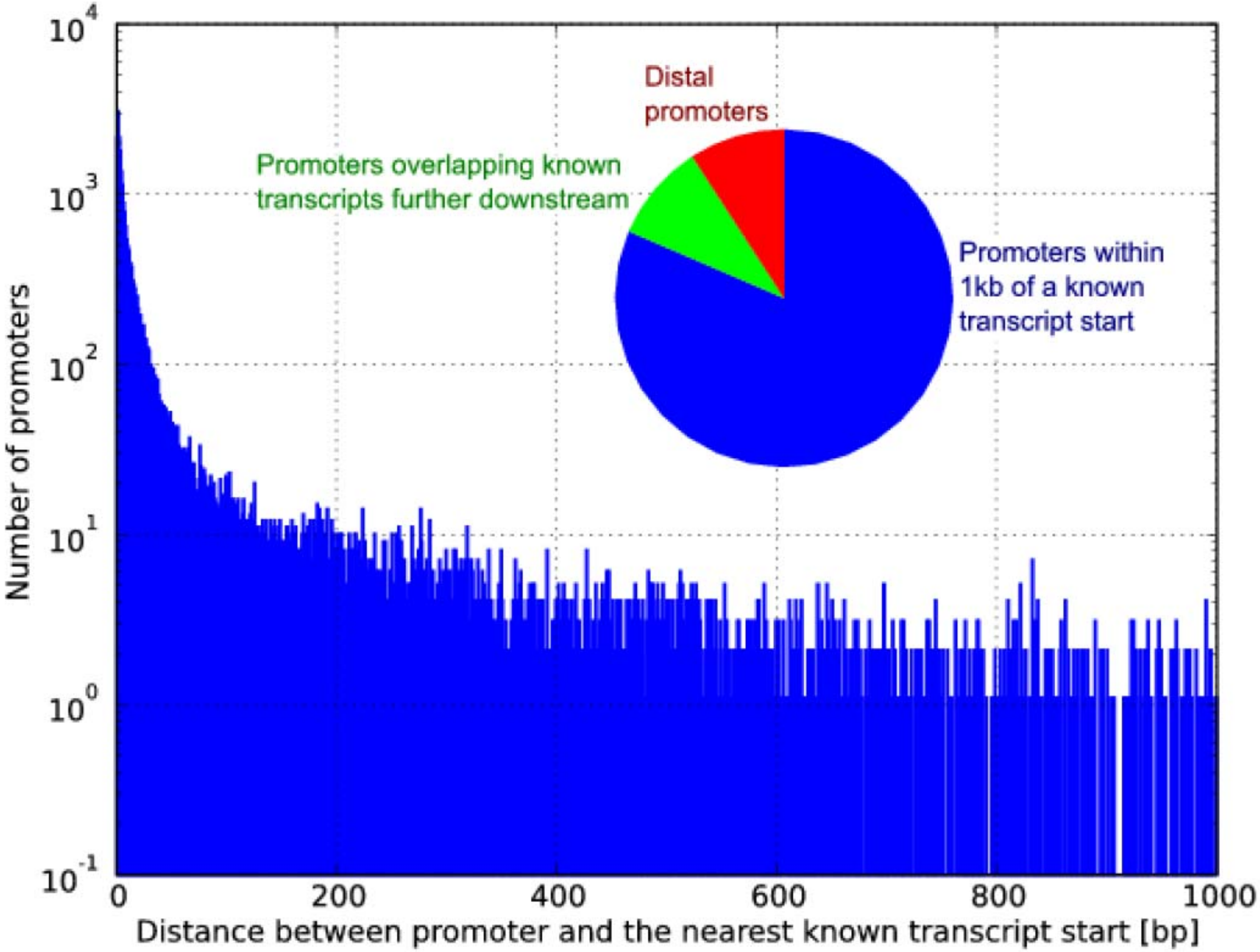
Supplementary Figure 2



Supplementary Figure 3

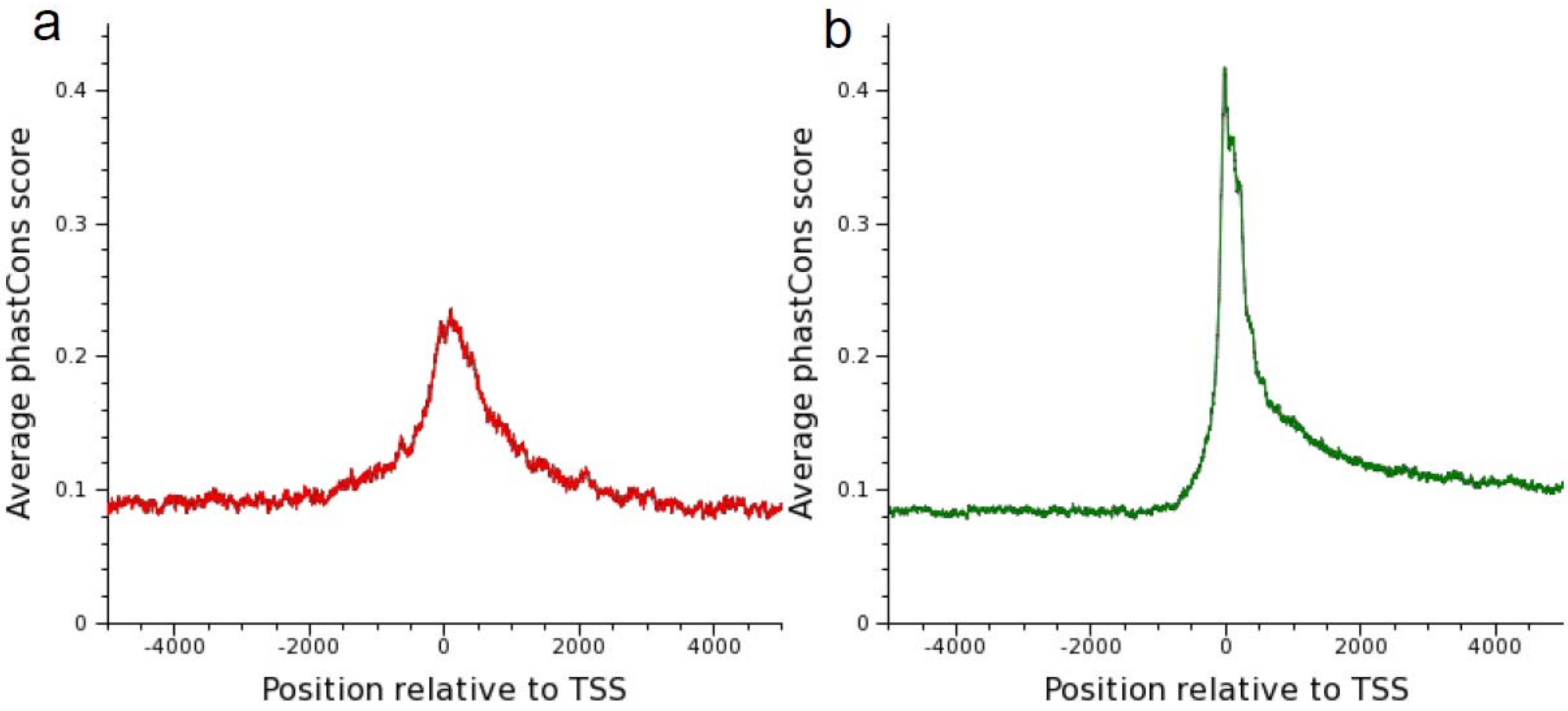


Supplementary Figure 4



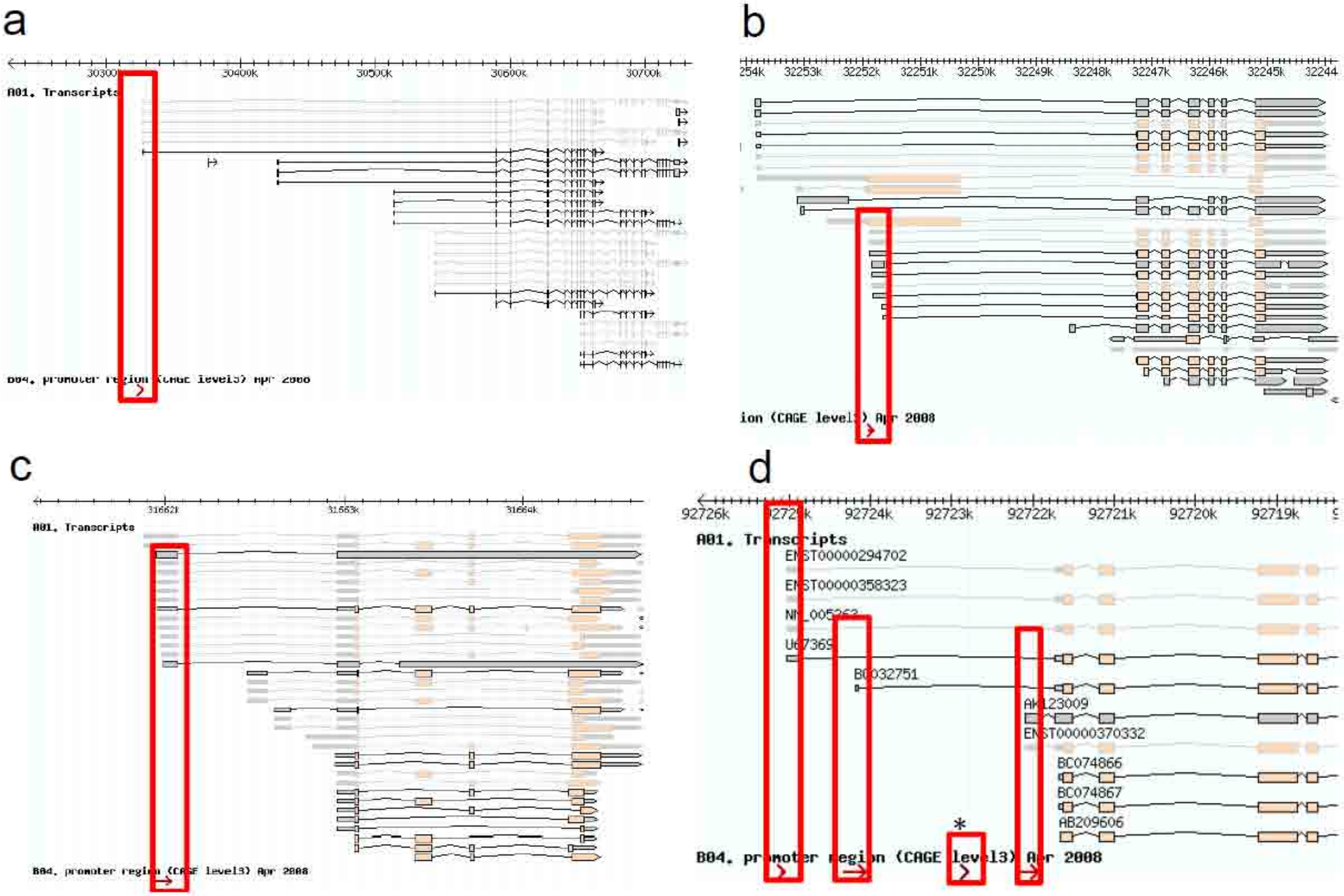
Nature Genetics: doi:10.1038/ng.375

Supplementary Figure 5



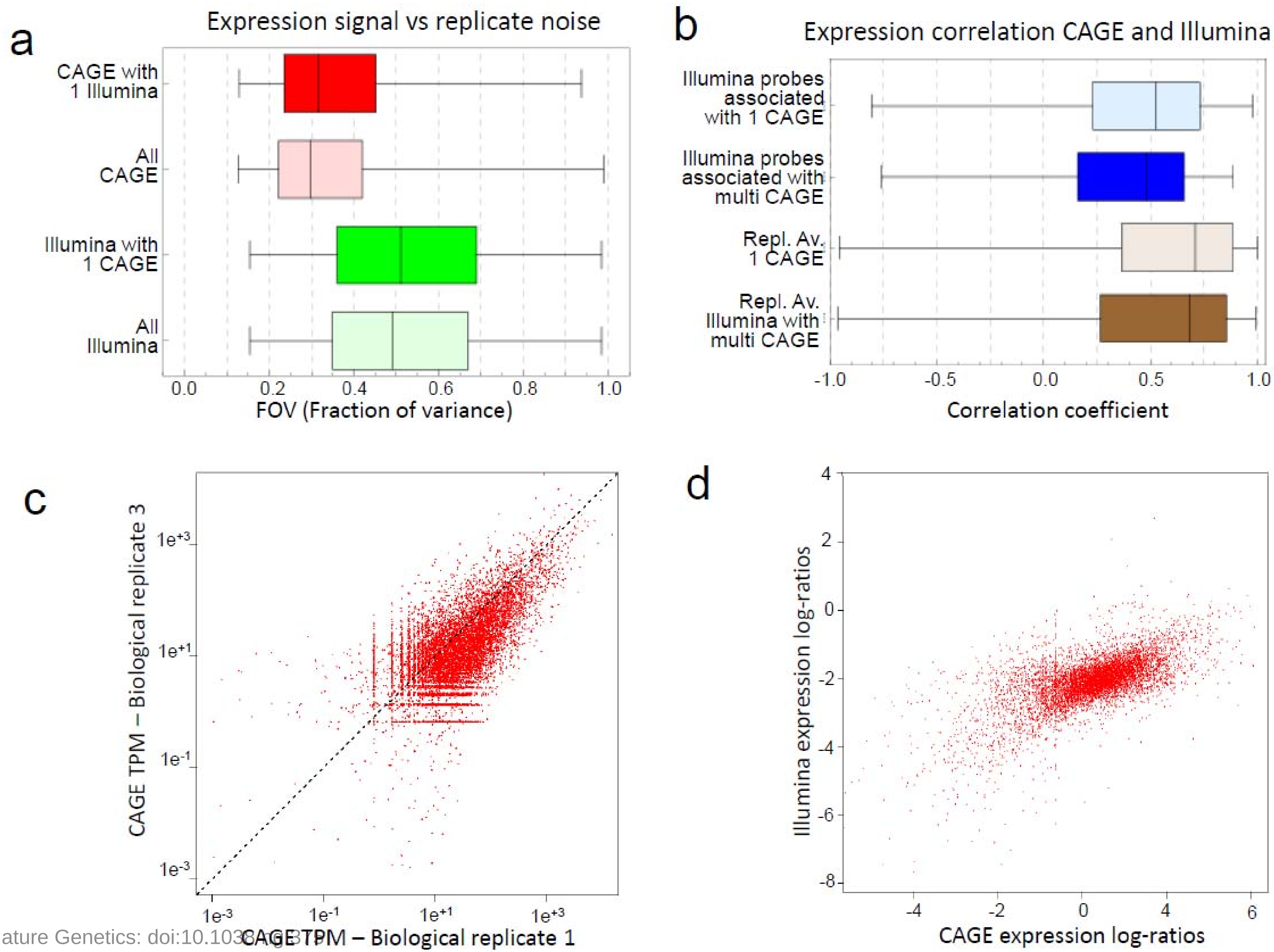
Nature Genetics: doi:10.1038/ng.375

Supplementary Figure 6

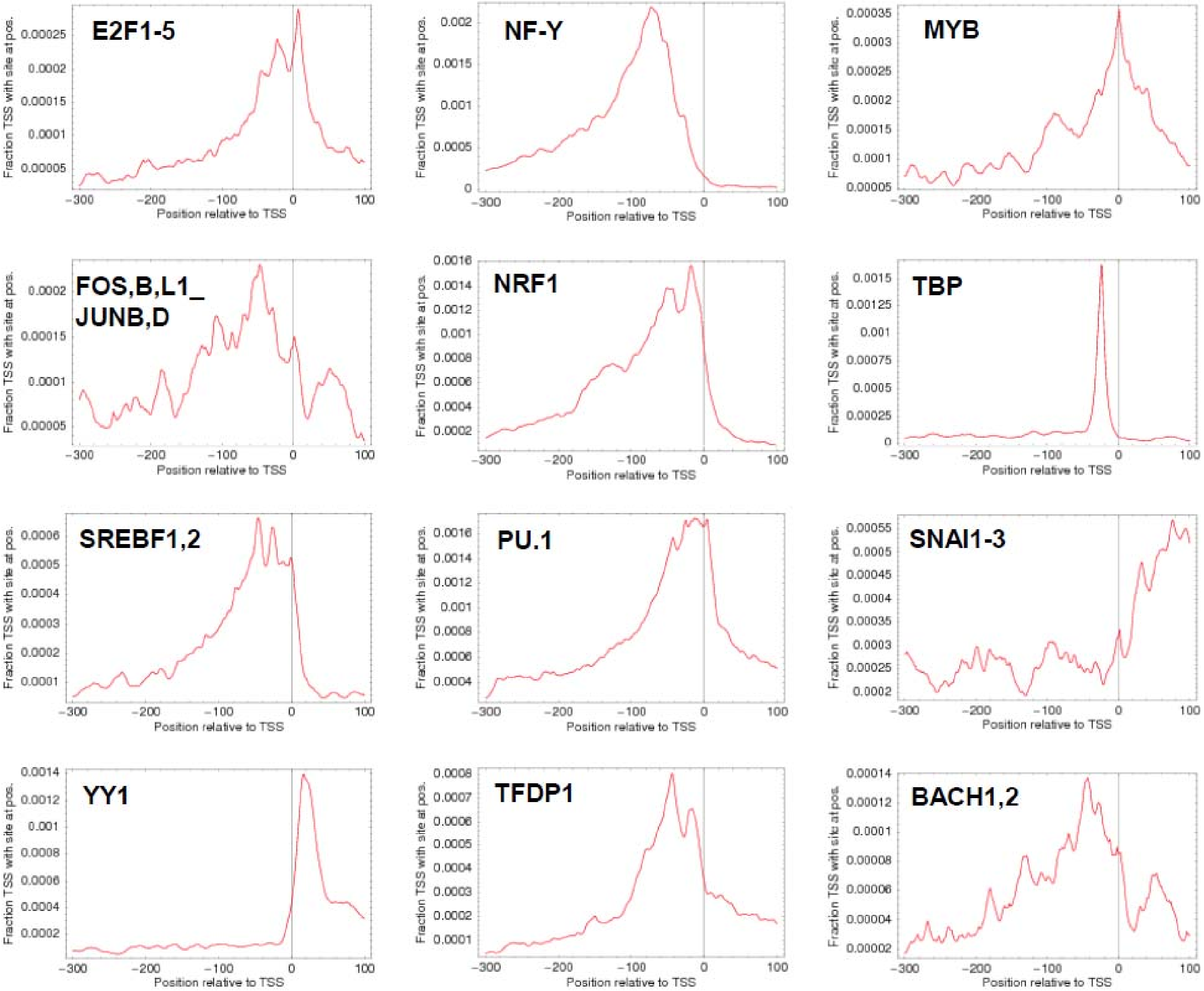


Nature Genetics: doi:10.1038/ng.375

Supplementary Figure 7

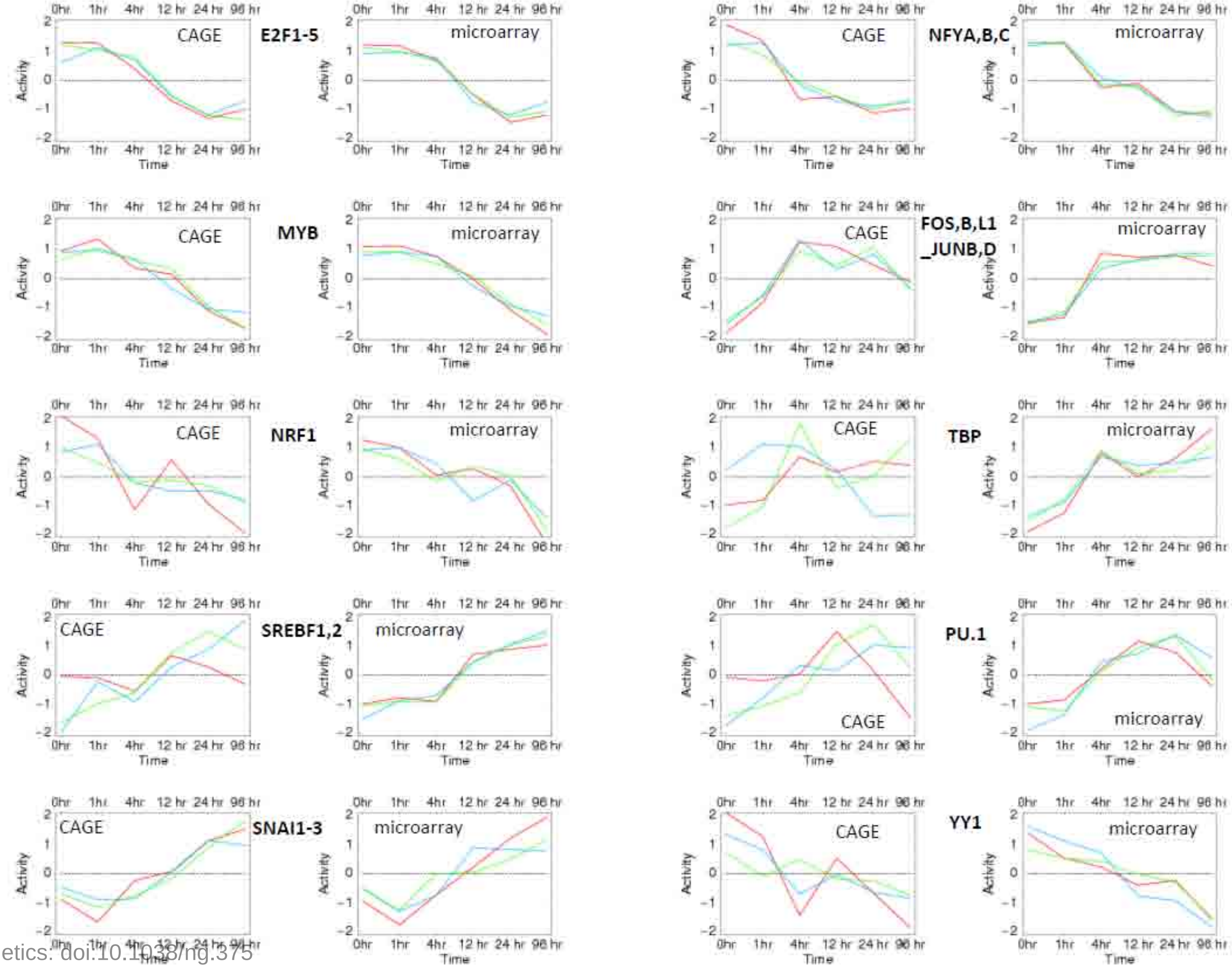


Supplementary Figure 8

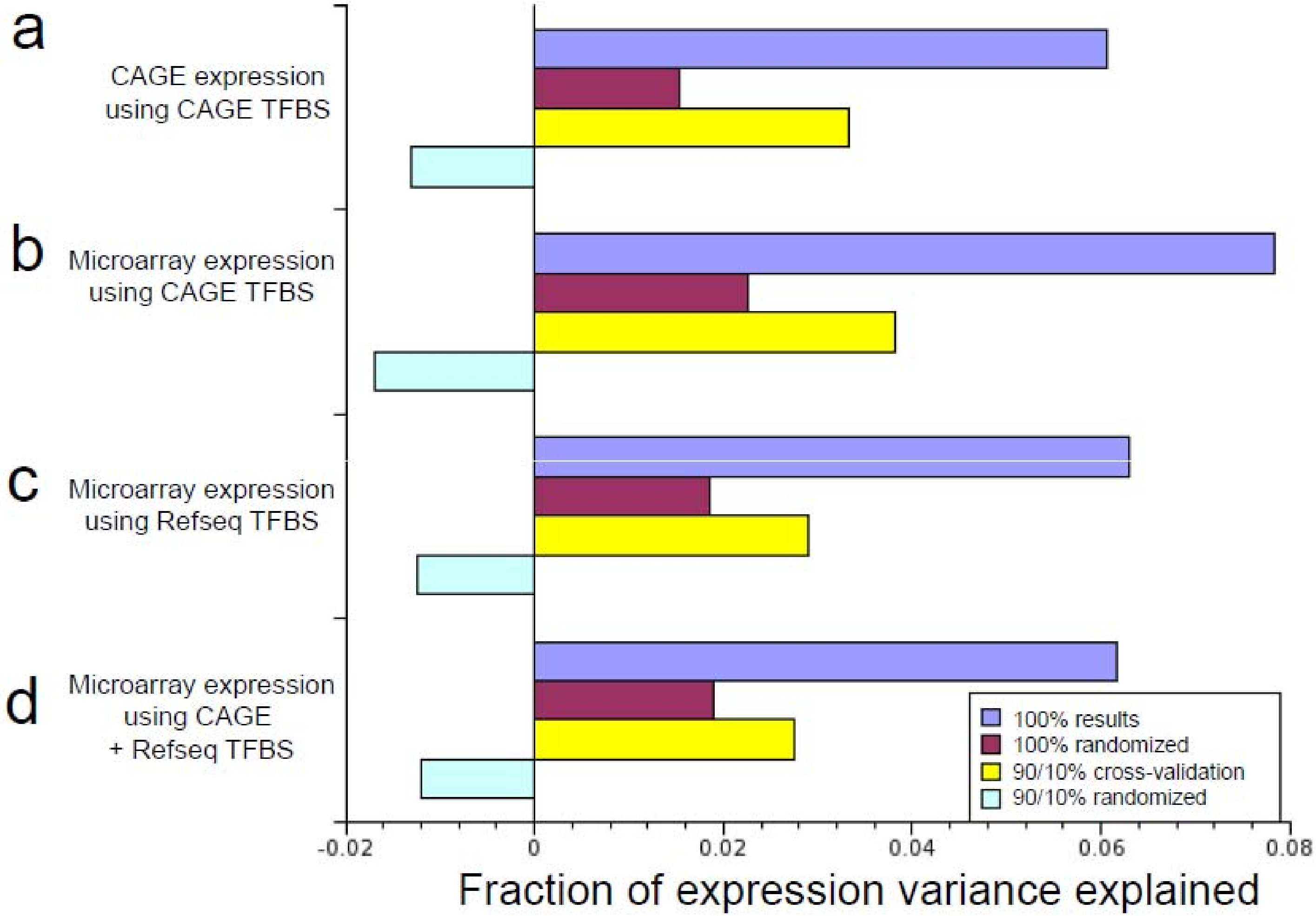


Nature Genetics: doi:10.1038/ng.375

Supplementary Figure 9

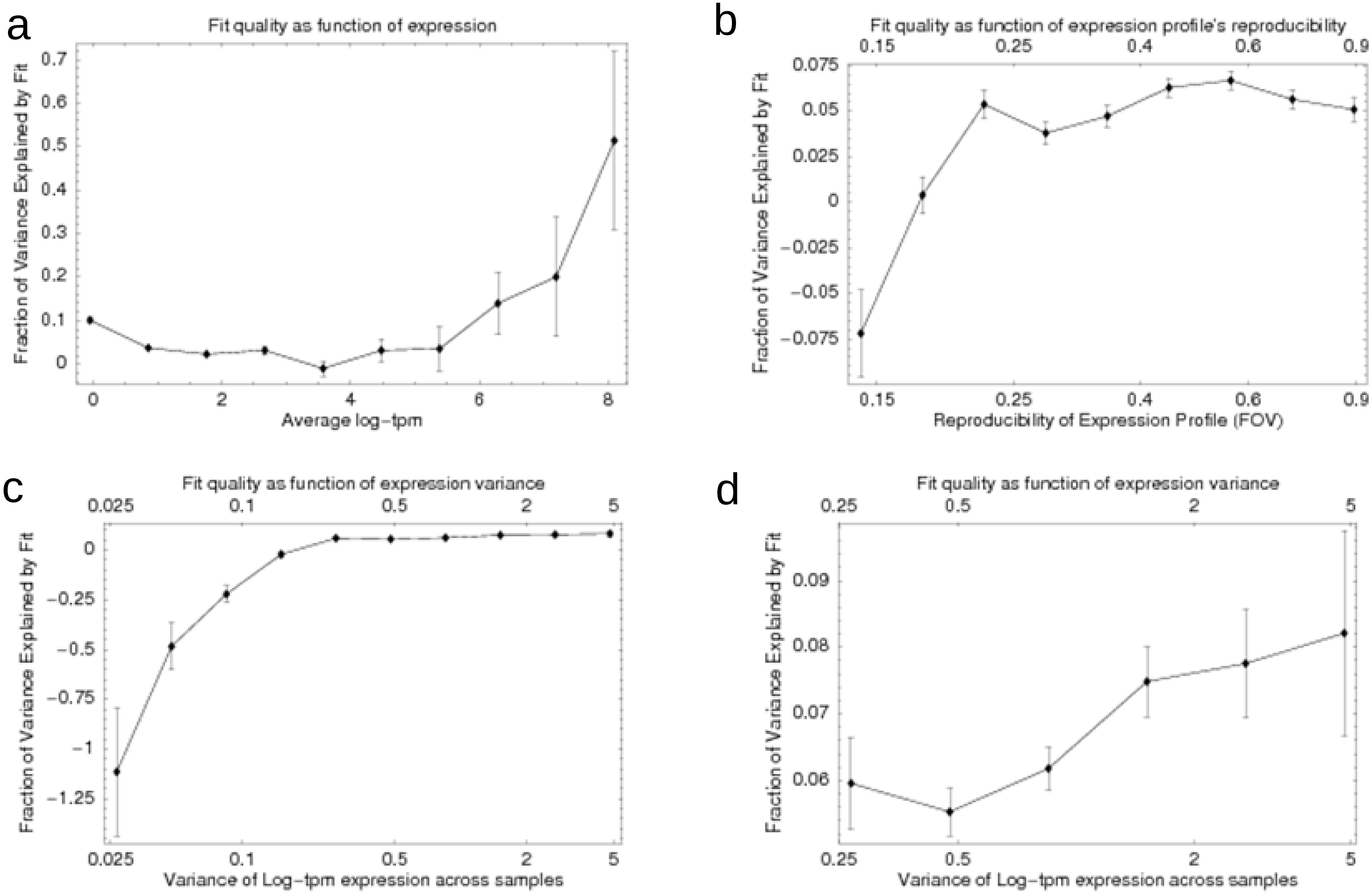


Supplementary Figure 10



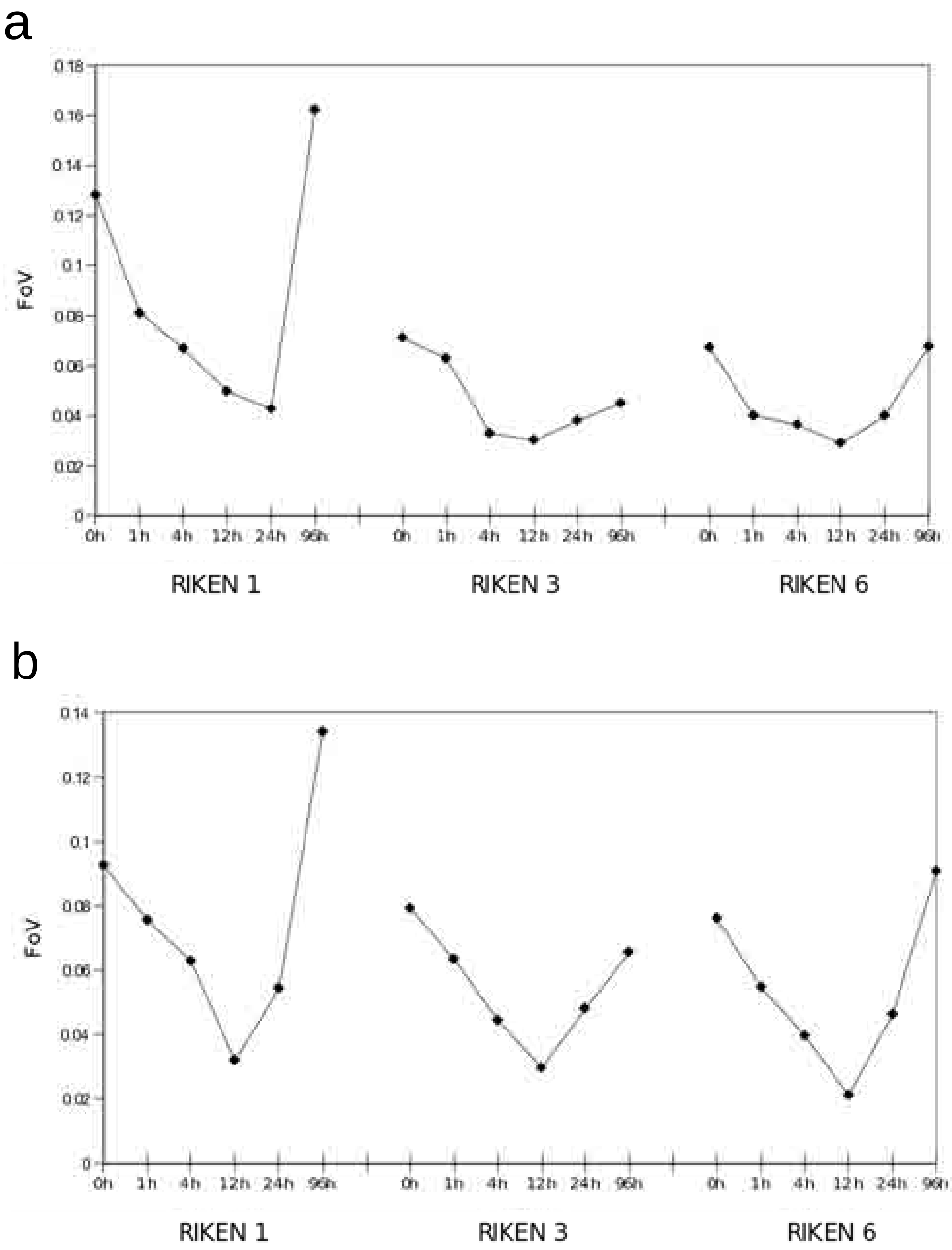
Nature Genetics: doi:10.1038/ng.375

Supplementary Figure 11



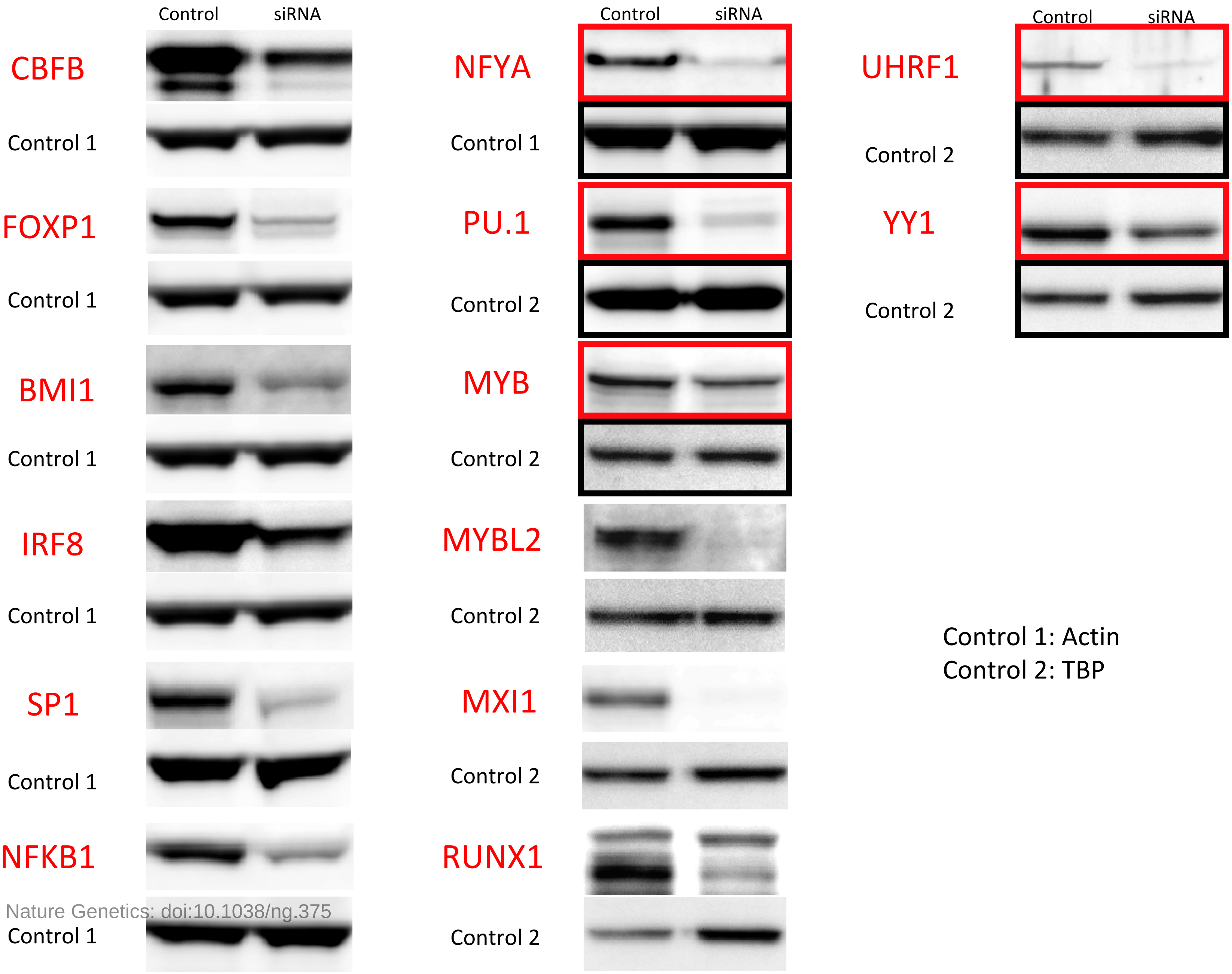
Nature Genetics: doi:10.1038/ng.375

Supplementary Figure 12



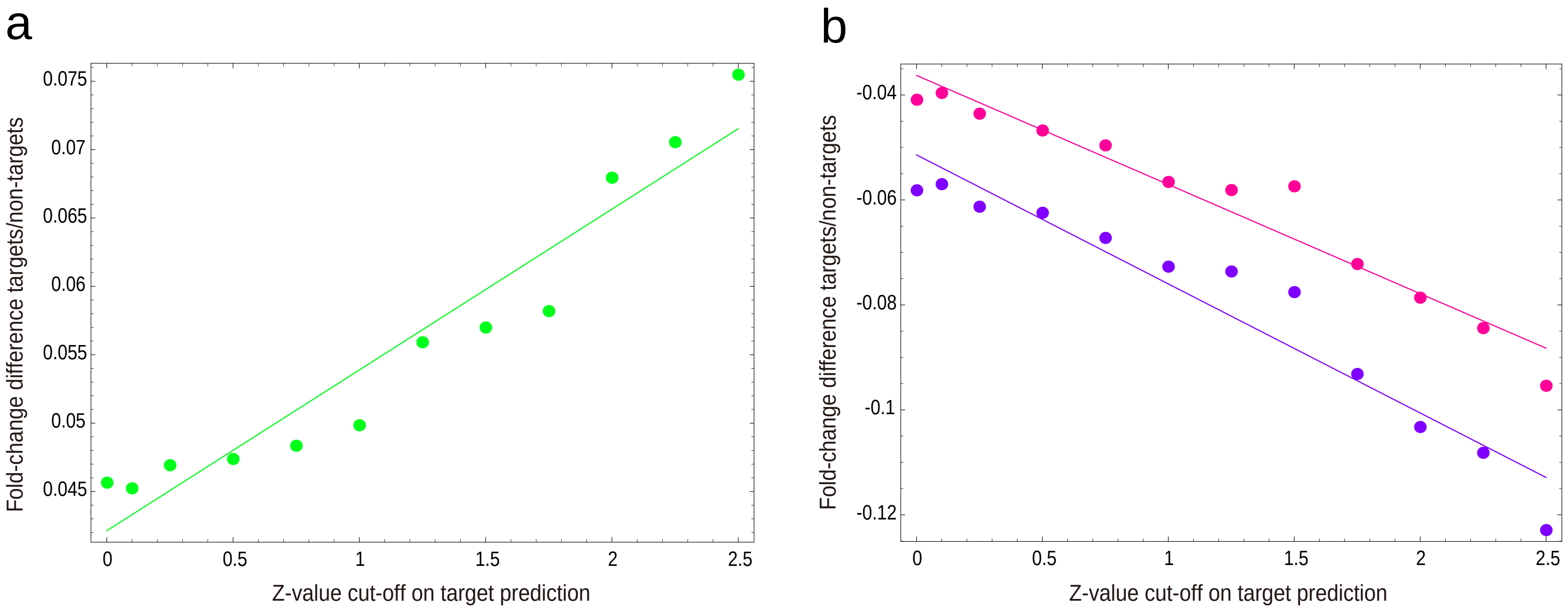
Nature Genetics: doi:10.1038/ng.375

Supplementary Figure 13

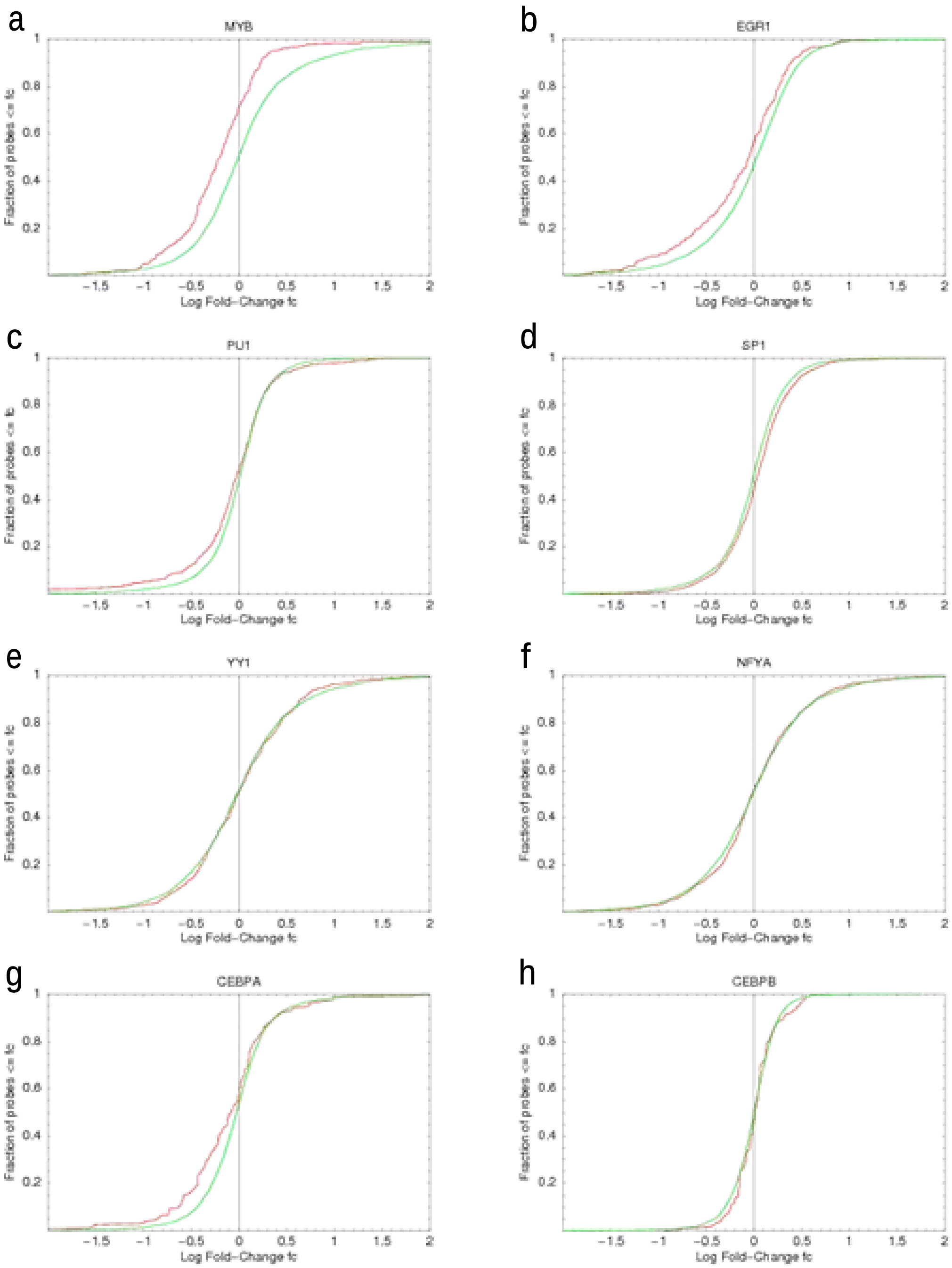


Control 1: Actin
Control 2: TBP

Supplementary Figure 14

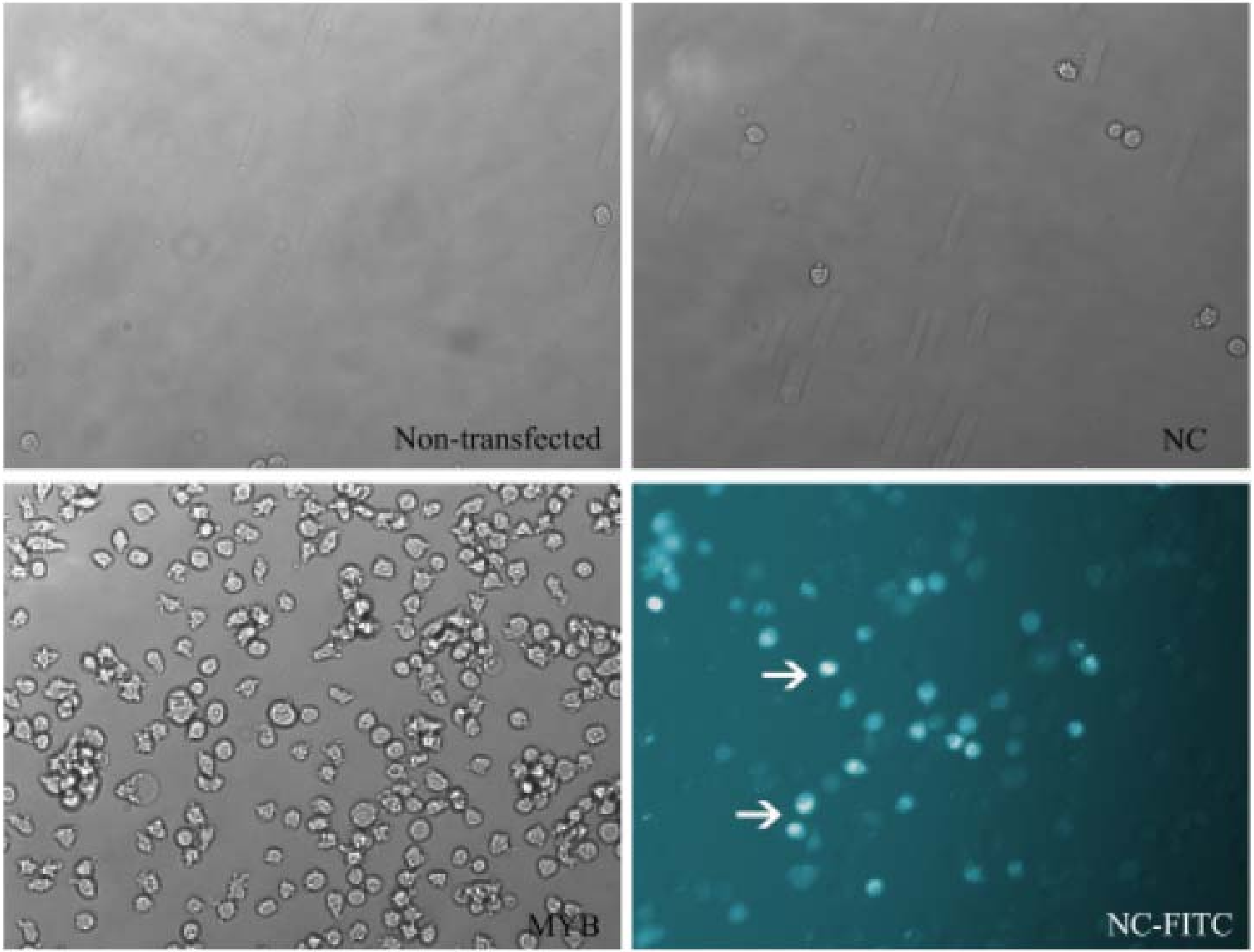


Supplementary Figure 15



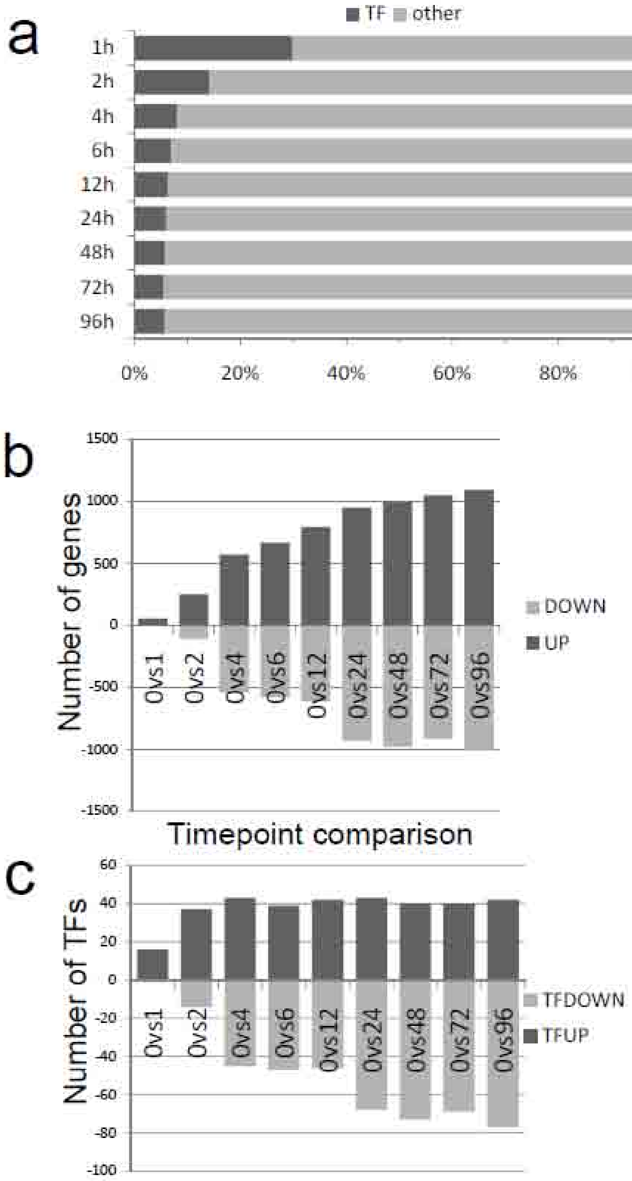
Nature Genetics: doi:10.1038/ng.375

Supplementary Figure 16



Nature Genetics: doi:10.1038/ng.375

Supplementary Figure 17



Supplementary Table 1 Distribution of the number of promoters per gene (zero counts not shown).

Number of promoters	Number of genes
1	3885
2	2176
3	1305
4	780
5	446
6	279
7	178
8	119
9	96
10	56
11	32
12	28
13	13
14	14
15	10
16	11
17	6
18	8
19	2
20	2
21	1
22	2
24	2
29	1
Total	9452

Note: We identified 9452 genes with at least one CAGE-defined promoter. The promoters shown in this table account for 24,327 out of the 29,857 promoters identified in total. 300 promoters are associated with two genes and 8 promoters with three genes. The remaining 5530 promoters were not assigned to any gene.

Supplementary Table 2 Distribution of the number of promoters per promoter region (zero counts not shown).

Number of promoters	Number of promoter regions
1	8600
2	2570
3	1388
4	784
5	459
6	280
7	179
8	112
9	79
10	55
11	26
12	23
13	13
14	7
15	10
16	3
17	6
18	6
19	2
21	2
27	1
42	1
46	1
Total	14607

Note: Promoter regions contain one or more CAGE defined promoters that are less than 400bp apart.

Supplementary Table 3 ChIP-chip enrichment in predicted targets compared to predicted non-targets**.**

% genes with CHIP signal		p-val	Enrichment	CHIP data
predicted targets	non targets		target/non	
16.8%	2.0%	2.00E-17	8.3	SRF[4]
35.9%	6.3%	6.60E-263	5.7	GABPA[5]
14.0%	5.4%	1.10E-28	2.6	YY1[6]
8.3%	3.7%	9.40E-07	2.3	RUNX1[5]
30.0%	14.3%	1.70E-67	2.1	ELF1[5]
55.6%	38.6%	3.50E-39	1.4	E2F4[4], E2F1,E2F4,E2F6[7]
43.7%	28.9%	1.10E-17	1.5	SP1 this publication
18.6%	15.9%	5.50E-25	1.2	MYC[8]
35.2%	24.7%	5.70E-04	1.4	PU.1 this publication
17.5%	14.0%	2.00E-02	1.2	ETS1[5]

Note: Public ChIP on chip data were extracted for SRF, E2F4, ELF1, ETS1, GABPA, RUNX1, YY1, E2F1, E2F4, E2F6, and MYC. All edges reported by the respective papers were converted into a matrix to Entrez geneID relationship (edge). These were then compared to the predictions and significance tested using Fisher's exact test. For the SP1 and PU.1 comparisons we used in house ChIP data from hybridizations on Affymetrix promoter tiling arrays (described in the **Methods**).

References

4. Cooper, S.J., Trinklein, N.D., Nguyen, L. & Myers, R.M. Serum response factor binding sites differ in three human cell types. Genome Res 17, 136-44 (2007).

5. Hollenhorst, P.C., Shah, A.A., Hopkins, C. & Graves, B.J. Genome-wide analyses reveal properties of redundant and specific promoter occupancy within the ETS gene family. Genes Dev 21, 1882-94 (2007).

6. Xi, H. et al. Analysis of overrepresented motifs in human core promoters reveals dual regulatory roles of YY1. Genome Res 17, 798-806 (2007).

7. Xu, X. et al. A comprehensive ChIP-chip analysis of E2F1, E2F4, and E2F6 in normal and tumor cells reveals interchangeable roles of E2F family members. Genome Res 17, 1550-61 (2007).

8. Zeller, K.I. et al. Global mapping of c-Myc binding sites and target gene networks in human B cells. Proc Natl Acad Sci U S A 103, 17834-9 (2006).

Supplementary Table 4 Significance and reproducibility of the 30 core motifs.

Motif	Significance (z-value)	Reproducibility (FOV)
E2F1-5	15.94	0.98
NFYA,B,C	14.35	0.96
MYB	9.77	0.98
FOS,B,L1_JUNB,D	8.66	0.92
NRF1	8.42	0.87
TBP	8.27	0.83
SREBF1,2	6.22	0.93
PU.1	6.07	0.78
SNAI1-3	5.59	0.93
YY1	5.4	0.83
TFDP1	5.35	0.9
BACH1,2	5.32	0.88
RUNX1-3	4.98	0.85
EBF1	4.79	0.85
NKX6-1,2	4.62	0.93
IRF1,2	4.59	0.83
ELK1,4_GABPA,B2	4.58	0.79
EGR1-3	4.52	0.82
ZIC1-3	4.4	0.96
NFATC1-3	4.38	0.9
FOXI1,J2	4.36	0.83
ATF5_CREB3	4.08	0.8
GATA4	4.04	0.9
TGIF1	4.01	0.86
RBPJ	3.92	0.8
POU6F1	3.91	0.91
SRF	3.88	0.83
TBX4,5	3.86	0.823
FOXO1,3,4	3.83	0.75
OCT4	3.79	0.85

Note: The significance is quantified by the overall z-value of the motif (activity relative to its standard-error) and the reproducibility is quantified by the fraction of variance that is reproduced across replicates and measurement technology, i.e. CAGE and microarray. Statistics for all other motifs are available from the FANTOM4 web resource.

Supplementary Table 5 Combined validation of the core network by publications, ChIP and siRNA.

source_matrix	Target_matrix	target_gene	zval	sirna bstat	chip	publications
NFYA,B,C	FOS,B,L1_JUNB,	JUNB	1.569594694			PMID: 11602259
NFATC1-3	NFATC1-3	NFATC1	7.574014304	not tested		PMID: 12121669
SRF	EGR1-3	EGR1	4.426911718	not tested	PMID: 17200232	PMID: 14769801
SRF	EGR1-3	EGR2	1.674251667	not tested	PMID: 17200232	PMID: 14769801
SRF	FOS,B,L1_JUNB,	FOSB	3.820039576	not tested	PMID: 17200232	PMID: 14769801
SRF	SRF	SRF	2.89399291	not tested	PMID: 17200232	PMID: 14769801
SRF	FOS,B,L1_JUNB,	FOSL1	2.645800288	not tested		PMID: 15806162
ELK1,4_GABPA,B	TBP	TBP	3.400760251	not tested	PMID: 17652178	PMID: 17074809
MYB	MYB	MYB	2.52919875	Auto not tested		PMID: 1944282
IRF1,2	IRF1,2	IRF2	7.657259227	not tested		PMID: 8106512
NFYA,B,C	SREBF1,2	SREBF2	7.204891704			PMID: 8900111
NFYA,B,C	E2F1-5	E2F1	5.67803211	2.868		PMID: 9218478 PMID: 12697671
TFDP1	MYB	MYB	8.173855601	not tested		PMID:10823896
E2F1-5	MYB	MYB	5.974570624			PMID:10823896
E2F1-5	E2F1-5	E2F1	7.414430094	Auto not tested		PMID:10823896 PMID: 10208422
ELK1,4_GABPA,B	EGR1-3	EGR1	1.85818518	not tested		PMID:11739517 PMID: 15449318
E2F1-5	E2F1-5	E2F2	3.305206423			PMID:11799067
FOS,B,L1_JUNB,D	FOS,B,L1_JUNB,	FOSL1	2.767910848	not tested		PMID:13679379
OCT4	ZIC1-3	ZIC2	2.9661653	not tested	PMID: 16153702	
SRF	EGR1-3	EGR3	4.580896367	not tested	PMID: 17200232	
RUNX1-3	RUNX1-3	RUNX1	2.98053755	Auto not tested	PMID: 17652178	
ELK1,4_GABPA,B	E2F1-5	E2F4	2.596023376	not tested	PMID: 17652178	
ELK1,4_GABPA,B	ELK1,4_GABPA,B	GABPA	2.220440851	not tested	PMID: 17652178	
ELK1,4_GABPA,B	ELK1,4_GABPA,B	GABPB2	4.977969572	not tested	PMID: 17652178	
E2F1-5	RUNX1-3	RUNX1	4.843770078		PMID: 17652178	
RUNX1-3	ZIC1-3	ZIC2	2.329344753		PMID: 17652178	
E2F1-5	EGR1-3	EGR3	4.836721441		PMID: 17908821	
E2F1-5	ELK1,4_GABPA,B	GABPB2	7.037663545		PMID: 17908821	
E2F1-5	NRF1	NRF1	1.8751406		PMID: 17908821	
E2F1-5	TFDP1	TFDP1	13.93723846		PMID: 17908821	
PU.1	NFYA,B,C	NFYC	6.70803037	0.994	this publication	
PU.1	RUNX1-3	RUNX1	2.123061788	3.132	this publication	

PU.1	ATF5_CREB3	CREB3	8.813642311		this publication	
PU.1	ELK1,4_GABPA,B	GABPA	3.774354352		this publication	
PU.1	ELK1,4_GABPA,B	GABPB2	2.239600599		this publication	
PU.1	IRF1,2	IRF1	2.229798497		this publication	
PU.1	IRF1,2	IRF2	2.796803675		this publication	
PU.1	RBPJ	RBPJ	2.365730014		this publication	
RUNX1-3	ELK1,4_GABPA,B	GABPB2	2.239905699	0.128		
SNAI1-3	PU.1	PU.1	4.024714295	0.978		
PU.1	ZIC1-3	ZIC2	1.976787975	1.664		
SNAI1-3	FOXO1,3,4	FOXO4	7.602870519	2.077		
SNAI1-3	SNAI1-3	SNAI3	5.679505428	3.12		
EGR1-3	SREBF1,2	SREBF2	2.241181679	5.445		
YY1	NFYA,B,C	NFYC	2.814834641	6.042		
NFYA,B,C	TBP	TBP	6.358809872	8.672		
SREBF1,2	ELK1,4_GABPA,B	ELK1	1.939428864	10.235		
YY1	MYB	MYB	11.02676842	11.877		
SNAI1-3	RUNX1-3	RUNX1	2.258065637	12.051		
EGR1-3	ELK1,4_GABPA,B	GABPA	3.289895083	12.202		
YY1	RBPJ	RBPJ	2.977633128	15.204		
RUNX1-3	RBPJ	RBPJ	1.959833645	20.13		
NFYA,B,C	RUNX1-3	RUNX1	4.125822185	20.168		
NFYA,B,C	NFYA,B,C	NFYB	2.874209569	21.805		
NFYA,B,C	E2F1-5	E2F4	3.742812048	48.663		
ATF5_CREB3	EGR1-3	EGR1	7.136748421	not tested		
ATF5_CREB3	EGR1-3	EGR2	1.694012401	not tested		
ATF5_CREB3	EGR1-3	EGR3	5.344340955	not tested		
ATF5_CREB3	ELK1,4_GABPA,B	ELK4	7.996073013	not tested		
ATF5_CREB3	FOS,B,L1_JUNB,	FOSB	4.508550527	not tested		
ATF5_CREB3	FOS,B,L1_JUNB,	FOSL1	6.860611945	not tested		
ATF5_CREB3	FOS,B,L1_JUNB,	JUND	2.509567527	not tested		
ATF5_CREB3	NFATC1-3	NFATC1	1.857010984	not tested		
ATF5_CREB3	NFYA,B,C	NFYC	1.773336841	not tested		
ATF5_CREB3	TFDP1	TFDP1	2.582213547	not tested		
ATF5_CREB3	YY1	YY1	4.004557219	not tested		
BACH1,2	E2F1-5	E2F3	3.406533193	not tested		

BACH1,2	ELK1,4_GABPA,B	GABPB2	1.968314104	not tested		
BACH1,2	FOS,B,L1_JUNB,	FOSL1	1.935474833	not tested		
BACH1,2	NFATC1-3	NFATC1	4.933620688	not tested		
BACH1,2	NFYA,B,C	NFYC	6.487565624	not tested		
BACH1,2	POU6F1	POU6F1	4.301253795	not tested		
BACH1,2	SREBF1,2	SREBF2	2.982922895	not tested		
EBF1	E2F1-5	E2F3	1.632179393	not tested		
EBF1	ELK1,4_GABPA,B	GABPB2	2.121984659	not tested		
ELK1,4_GABPA,B	EGR1-3	EGR3	3.078899814	not tested		
ELK1,4_GABPA,B	ELK1,4_GABPA,B	ELK4	4.127635881	not tested		
ELK1,4_GABPA,B	RBPJ	RBPJ	5.465343651	not tested		
ELK1,4_GABPA,B	RUNX1-3	RUNX1	6.885756512	not tested		
ELK1,4_GABPA,B	YY1	YY1	5.574669474	not tested		
FOS,B,L1_JUNB,D	ELK1,4_GABPA,B	GABPA	3.589622622	not tested		
FOS,B,L1_JUNB,D	ELK1,4_GABPA,B	GABPB2	2.054439753	not tested		
FOS,B,L1_JUNB,D	IRF1,2	IRF2	2.046561795	not tested		
FOS,B,L1_JUNB,D	NFYA,B,C	NFYC	9.408043774	not tested		
FOS,B,L1_JUNB,D	RUNX1-3	RUNX1	2.91208542	not tested		
FOXI1,J2	EGR1-3	EGR3	2.41650624	not tested		
FOXI1,J2	ELK1,4_GABPA,B	GABPB2	2.440238549	not tested		
FOXI1,J2	FOS,B,L1_JUNB,	FOSB	2.079930775	not tested		
FOXI1,J2	FOXO1,3,4	FOXO3	4.129701308	not tested		
FOXI1,J2	MYB	MYB	10.52829821	not tested		
FOXI1,J2	NFATC1-3	NFATC1	4.459770364	not tested		
FOXI1,J2	RUNX1-3	RUNX1	4.099492376	not tested		
FOXI1,J2	YY1	YY1	1.965387453	not tested		
FOXI1,J2	ZIC1-3	ZIC2	2.671543761	not tested		
FOXO1,3,4	ELK1,4_GABPA,B	GABPB2	2.946874089	not tested		
FOXO1,3,4	FOS,B,L1_JUNB,	JUND	3.670598151	not tested		
FOXO1,3,4	FOXO1,3,4	FOXO3	3.745552061	not tested		
FOXO1,3,4	FOXO1,3,4	FOXO4	2.341504494	not tested		
FOXO1,3,4	IRF1,2	IRF2	5.083448579	not tested		
FOXO1,3,4	NFATC1-3	NFATC1	2.119526355	not tested		
FOXO1,3,4	SREBF1,2	SREBF2	2.358907734	not tested		
GATA4	EGR1-3	EGR3	3.125234841	not tested		

GATA4	ELK1,4_GABPA,B	GABPB2	1.744594602	not tested		
GATA4	MYB	MYB	8.258204267	not tested		
GATA4	RUNX1-3	RUNX1	3.481500284	not tested		
GATA4	TBX4,5	TBX4	4.401821393	not tested		
IRF1,2	ELK1,4_GABPA,B	GABPB2	2.630526063	not tested		
IRF1,2	FOXO1,3,4	FOXO4	7.050856678	not tested		
NFATC1-3	E2F1-5	E2F3	3.405362148	not tested		
NFATC1-3	ELK1,4_GABPA,B	GABPA	1.97160488	not tested		
NFATC1-3	ELK1,4_GABPA,B	GABPB2	2.142735673	not tested		
NFATC1-3	FOXO1,3,4	FOXO1	3.295169501	not tested		
NFATC1-3	FOXO1,3,4	FOXO4	2.546652925	not tested		
NFATC1-3	IRF1,2	IRF2	1.520125445	not tested		
NFATC1-3	NFYA,B,C	NFYC	3.009714498	not tested		
NFATC1-3	RBPJ	RBPJ	1.674188877	not tested		
NKX6-1,2	FOXO1,3,4	FOXO1	3.543139119	not tested		
NKX6-1,2	FOXO1,3,4	FOXO4	5.148573798	not tested		
NKX6-1,2	NFYA,B,C	NFYC	3.030841658	not tested		
NKX6-1,2	RUNX1-3	RUNX1	3.427604325	not tested		
NRF1	E2F1-5	E2F1	2.618060994	not tested		
NRF1	E2F1-5	E2F4	3.350128475	not tested		
NRF1	ELK1,4_GABPA,B	GABPA	1.674342482	not tested		
NRF1	FOXO1,3,4	FOXO3	2.824834281	not tested		
NRF1	MYB	MYB	11.2302512	not tested		
NRF1	NFYA,B,C	NFYA	1.822572088	not tested		
NRF1	RBPJ	RBPJ	3.518785249	not tested		
NRF1	SREBF1,2	SREBF2	2.721535511	not tested		
NRF1	TFDP1	TFDP1	6.09066965	not tested		
NRF1	ZIC1-3	ZIC2	7.306767562	not tested		
OCT4	FOS,B,L1_JUNB,	JUND	3.493216149	not tested		
OCT4	MYB	MYB	3.579155704	not tested		
OCT4	NFYA,B,C	NFYC	3.71271422	not tested		
OCT4	RBPJ	RBPJ	2.402570056	not tested		
OCT4	RUNX1-3	RUNX1	3.545673719	not tested		
OCT4	SREBF1,2	SREBF2	4.374562652	not tested		
OCT4	SRF	SRF	4.824099614	not tested		

POU6F1	E2F1-5	E2F3	2.685797948	not tested		
POU6F1	ELK1,4_GABPA,B	GABPB2	1.755568671	not tested		
POU6F1	POU6F1	POU6F1	4.660893385	not tested		
RBPJ	ELK1,4_GABPA,B	GABPB2	2.165787396	not tested		
RBPJ	FOXO1,3,4	FOXO4	6.097810076	not tested		
RBPJ	IRF1,2	IRF2	5.076706077	not tested		
SRF	FOS,B,L1_JUNB,	JUND	5.783135674	not tested		
TBP	E2F1-5	E2F3	2.857393728	not tested		
TBP	FOS,B,L1_JUNB,	FOSL1	2.425948966	not tested		
TBP	FOXO1,3,4	FOXO4	2.612169141	not tested		
TBP	NFATC1-3	NFATC1	8.97703999	not tested		
TBP	ZIC1-3	ZIC2	2.111573809	not tested		
TBX4,5	ELK1,4_GABPA,B	GABPA	1.670149731	not tested		
TBX4,5	NFYA,B,C	NFYB	5.964359574	not tested		
TBX4,5	RUNX1-3	RUNX2	4.234837742	not tested		
TBX4,5	SNAI1-3	SNAI1	1.520088468	not tested		
TBX4,5	ZIC1-3	ZIC2	8.226256346	not tested		
TFDP1	E2F1-5	E2F4	3.901491318	not tested		
TFDP1	EGR1-3	EGR3	3.707947747	not tested		
TFDP1	ELK1,4_GABPA,B	GABPB2	6.162274316	not tested		
TFDP1	FOS,B,L1_JUNB,	FOSB	5.059060228	not tested		
TFDP1	FOXO1,3,4	FOXO3	1.878408879	not tested		
TFDP1	RUNX1-3	RUNX1	3.791060632	not tested		
TFDP1	RUNX1-3	RUNX2	2.680263053	not tested		
TFDP1	SREBF1,2	SREBF2	6.949179577	not tested		
TFDP1	ZIC1-3	ZIC2	13.03783664	not tested		
TGIF1	E2F1-5	E2F3	1.818890694	not tested		
TGIF1	RBPJ	RBPJ	1.811678565	not tested		
TGIF1	SREBF1,2	SREBF2	2.980691965	not tested		
TGIF1	TGIF1	TGIF1	2.934571213	not tested		
ZIC1-3	ELK1,4_GABPA,B	GABPB2	1.784772156	not tested		
ZIC1-3	FOXO1,3,4	FOXO1	3.592655854	not tested		
ZIC1-3	FOXO1,3,4	FOXO4	6.993507354	not tested		
ZIC1-3	TGIF1	TGIF1	3.406737198	not tested		
E2F1-5	RBPJ	RBPJ	3.800367952			

E2F1-5	TBX4,5	TBX4	3.635988546			
EGR1-3	NFATC1-3	NFATC3	2.578611777			
EGR1-3	RUNX1-3	RUNX1	2.666920831			
EGR1-3	TGIF1	TGIF1	5.117747314			
EGR1-3	ZIC1-3	ZIC2	2.637833205			
MYB	ELK1,4_GABPA,B	GABPA	2.361050583			
MYB	ELK1,4_GABPA,B	GABPB2	4.015416304			
MYB	NFYA,B,C	NFYA	3.558187102			
MYB	RBPJ	RBPJ	5.010863247			
MYB	RUNX1-3	RUNX2	4.457884691			
NFYA,B,C	E2F1-5	E2F2	3.708968262			
NFYA,B,C	SREBF1,2	SREBF1	8.857573327			
NFYA,B,C	SRF	SRF	3.71119832			
NFYA,B,C	ZIC1-3	ZIC2	2.218719109			
PU.1	E2F1-5	E2F4	2.114817748			
PU.1	FOXO1,3,4	FOXO4	3.849981777			
RUNX1-3	SREBF1,2	SREBF2	2.519341703			
SNAI1-3	IRF1,2	IRF2	5.614238481			
SREBF1,2	ELK1,4_GABPA,B	GABPB2	2.341530547			
SREBF1,2	FOXO1,3,4	FOXO1	2.600580031			
SREBF1,2	NFATC1-3	NFATC3	2.795295314			
YY1	EGR1-3	EGR3	2.102396029			
YY1	ELK1,4_GABPA,B	GABPA	1.579162347			
YY1	ELK1,4_GABPA,B	GABPB2	4.399538622			
YY1	FOS,B,L1_JUNB,	JUND	3.171430884			
YY1	RUNX1-3	RUNX1	2.955544083			
YY1	SREBF1,2	SREBF2	7.029712983			

Supplementary Table 6 Gene Ontology terms enriched in predicted target genes of core 30 motifs.

CAGE

SREBF1,2

GO	GO as name	P-Value
GO:000577	vacuole	8.21E-14
GO:004443	vacuolar part	1.14E-09
GO:000577	vacuolar membrane	1.14E-09
GO:000576	lysosome	1.91E-09
GO:000032	lytic vacuole	1.91E-09
GO:001982	cation-transporting ATPase activity	3.91E-06
GO:004696	hydrogen ion transporting ATPase activit	9.83E-06
GO:001599	proton transport	1.52E-05
GO:000681	hydrogen transport	1.72E-05
GO:003141	cytoplasmic vesicle	1.95E-05

NFATC1-3

GO	GO as name	P-Value
GO:002261	biological adhesion	3.22E-10
GO:000715	cell adhesion	3.22E-10
GO:000695	immune response	1.73E-09
GO:000588	plasma membrane	6.84E-09
GO:000960	response to external stimulus	8.53E-08
GO:000237	immune system process	2.23E-07
GO:003250	multicellular organismal process	1.17E-05
GO:003122	intrinsic to plasma membrane	1.27E-05
GO:000561	extracellular space	2.56E-05
GO:000487	receptor activity	2.57E-05

NRF1

GO	GO as name	P-Value
GO:000563	nucleus	5.04E-13
GO:004442	nuclear part	2.03E-09
GO:000838	RNA splicing	2.03E-07
GO:000639	mRNA processing	4.78E-07
GO:000625	DNA metabolic process	6.19E-07
GO:001602	membrane	-6.19E-07

Illumina

SREBF1,2

GO	GO as name	P-Value
GO:000577	vacuole	1.71E-31
GO:000576	lysosome	6.20E-12
GO:000032	lytic vacuole	6.20E-12
GO:004443	vacuolar part	5.96E-11
GO:000577	vacuolar membrane	5.96E-11
GO:000576	lysosomal membrane	1.96E-07
GO:003141	cytoplasmic vesicle	1.96E-07
GO:000576	endosome	2.91E-07
GO:003198	vesicle	3.67E-07
GO:000577	late endosome	2.11E-06

NFATC1-3

GO	GO as name	P-Value
GO:000561	extracellular space	2.26E-14
GO:004442	extracellular region part	2.18E-10
GO:000960	response to external stimulus	5.98E-10
GO:000961	response to wounding	1.36E-09
GO:000695	immune response	1.22E-08
GO:000237	immune system process	1.67E-08
GO:000715	cell communication	6.14E-08
GO:000695	inflammatory response	8.12E-08
GO:000716	signal transduction	3.14E-07
GO:000510	receptor binding	4.28E-07

NRF1

GO	GO as name	P-Value
GO:000563	nucleus	8.34E-23
GO:004442	nuclear part	3.59E-14
GO:000372	RNA binding	4.95E-13
GO:004323	intracellular membrane-bound organell	1.64E-12
GO:004322	membrane-bound organelle	1.64E-12
GO:001607	mRNA metabolic process	2.20E-12

GO:000639	RNA processing	6.19E-07
GO:000372	RNA binding	1.07E-06
GO:000569	chromosome	2.42E-06
GO:001607	mRNA metabolic process	2.42E-06

TFDP1

GO	GO as name	P-Value
GO:000563	nucleus	7.15E-14
GO:000613	nucleobase, nucleoside, nucleotide and	8.75E-07
GO:000625	DNA metabolic process	1.00E-06
GO:000367	DNA binding	1.39E-05
GO:000551	protein binding	2.89E-05
GO:004442	nuclear part	2.89E-05
GO:000367	nucleic acid binding	3.76E-05
GO:003132	regulation of cellular metabolic process	6.33E-05
GO:001921	regulation of nucleobase, nucleoside, nu	6.77E-05
GO:000635	regulation of transcription from RNA pol	6.77E-05

RUNX1-3

No GO term enrichment

E2F1-5

GO	GO as name	P-Value
GO:000625	DNA metabolic process	1.02E-47
GO:000569	chromosome	1.27E-25
GO:004442	chromosomal part	3.95E-25
GO:000563	nucleus	4.35E-24
GO:000704	cell cycle	1.70E-22
GO:000613	nucleobase, nucleoside, nucleotide and	1.94E-22
GO:000971	response to endogenous stimulus	4.80E-21

GO:000639	mRNA processing	1.14E-11
GO:000367	nucleic acid binding	1.14E-11
GO:000639	RNA processing	1.37E-11
GO:000562	intracellular	1.31E-10

TFDP1

GO	GO as name	P-Value
GO:000563	nucleus	2.11E-19
GO:000613	nucleobase, nucleoside, nucleotide and	8.12E-15
GO:000625	DNA metabolic process	1.90E-14
GO:005127	chromosome organization and biogenes	1.57E-11
GO:004442	nuclear part	2.27E-10
GO:004322	intracellular organelle	7.37E-10
GO:004322	organelle	7.37E-10
GO:004323	intracellular membrane-bound organelle	9.34E-10
GO:004322	membrane-bound organelle	9.34E-10
GO:000632	establishment and/or maintenance of ch	2.68E-09

RUNX1-3

GO	GO as name	P-Value
GO:000237	immune system process	1.12E-08
GO:004442	extracellular region part	0.00011
GO:000960	response to external stimulus	0.00011
GO:000510	receptor binding	0.000599
GO:004873	multicellular organismal development#s	0.000599
GO:000561	extracellular space	0.00077
GO:000695	immune response	0.000866
GO:000961	response to wounding	0.00102
GO:004851	multicellular organismal development#s	0.00111
GO:000695	defense response	0.00219

E2F1-5

GO	GO as name	P-Value
GO:000625	DNA metabolic process	8.62E-39
GO:000563	nucleus	1.60E-28
GO:004442	chromosomal part	2.81E-28
GO:000569	chromosome	9.16E-27
GO:000704	cell cycle	7.60E-24
GO:002240	cell cycle phase	1.21E-21
GO:000613	nucleobase, nucleoside, nucleotide and	1.30E-20

GO:005127	chromosome organization and biogenes	2.66E-20
GO:002240	cell cycle phase	4.44E-20
GO:000697	response to DNA damage stimulus	3.51E-19

ELK1,4_GABPA,B2

GO	GO as name	P-Value
GO:000563	nucleus	6.30E-14
GO:004442	nuclear part	8.44E-11
GO:000639	RNA processing	5.11E-10
GO:000613	nucleobase, nucleoside, nucleotide and	5.11E-10
GO:001046	gene expression	5.41E-09
GO:001607	RNA metabolic process	7.93E-09
GO:000367	nucleic acid binding	4.89E-08
GO:000639	mRNA processing	7.77E-08
GO:000838	RNA splicing	1.25E-07
GO:004328	biopolymer metabolic process	1.32E-07

FOXI1,J2

GO	GO as name	P-Value
GO:000370	transcription factor activity	1.83E-10
GO:000635	regulation of transcription, DNA-depend	4.99E-08
GO:004544	regulation of transcription	4.99E-08
GO:000635	transcription, DNA-dependent	1.63E-07
GO:003277	RNA biosynthetic process	1.63E-07
GO:001921	regulation of nucleobase, nucleoside, nu	1.68E-07
GO:001046	gene expression#regulation of gene expr	1.75E-07
GO:000635	transcription	3.32E-07
GO:003132	regulation of cellular metabolic process	4.26E-07
GO:000367	DNA binding	1.04E-06

IRF1,2

GO	GO as name	P-Value
GO:000695	defense response	6.91E-03
GO:000237	immune system process	6.91E-03
GO:000504	scavenger receptor activity	8.58E-03

GO:005127	chromosome organization and biogenes	8.86E-20
GO:000027	mitotic cell cycle	1.49E-19
GO:002240	cell cycle process	3.76E-17

ELK1,4_GABPA,B2

GO	GO as name	P-Value
GO:000563	nucleus	9.15E-31
GO:004442	nuclear part	6.48E-25
GO:000639	RNA processing	3.53E-24
GO:001607	mRNA metabolic process	8.01E-23
GO:000639	mRNA processing	6.09E-22
GO:000838	RNA splicing	1.36E-20
GO:000613	nucleobase, nucleoside, nucleotide and	2.02E-20
GO:001046	gene expression	3.61E-20
GO:001607	RNA metabolic process	4.94E-20
GO:004322	membrane-bound organelle	7.56E-20

FOXI1,J2

GO	GO as name	P-Value
GO:000635	regulation of transcription, DNA-depend	3.54E-08
GO:004544	regulation of transcription	3.54E-08
GO:001921	regulation of nucleobase, nucleoside, nu	9.91E-08
GO:000635	transcription, DNA-dependent	9.91E-08
GO:003277	RNA biosynthetic process	9.91E-08
GO:000370	transcription factor activity	9.91E-08
GO:000635	transcription	9.91E-08
GO:003132	regulation of cellular metabolic process	1.92E-07
GO:001046	gene expression#regulation of gene expr	2.36E-07
GO:001922	metabolic process#regulation of metabo	3.55E-06

IRF1,2

GO	GO as name	P-Value
GO:000237	immune system process	1.22E-05
GO:000695	defense response	2.95E-05
GO:004442	extracellular region part	2.95E-05
GO:000695	immune response	3.62E-05
GO:000504	scavenger receptor activity	0.000122
GO:000561	extracellular space	0.000122
GO:000510	receptor binding	0.000885
GO:002261	biological adhesion	0.000885

NFYA,B,C

GO	GO as name	P-Value
GO:004442	chromosomal part	1.52E-25
GO:000078	chromatin	1.52E-25
GO:000569	chromosome	4.20E-23
GO:000625	DNA metabolic process	1.33E-22
GO:005127	chromosome organization and biogenes	3.18E-21
GO:000563	nucleus	1.02E-17
GO:000632	DNA packaging	5.24E-17
GO:000632	establishment and/or maintenance of ch	1.34E-16
GO:000704	cell cycle	6.17E-16
GO:000633	chromatin assembly or disassembly	1.80E-15

SNAI1-3

GO	GO as name	P-Value
GO:001602	membrane	5.65E-06
GO:000588	integral to plasma membrane	1.41E-05
GO:000588	plasma membrane	1.41E-05
GO:003122	intrinsic to plasma membrane	1.41E-05
GO:004445	plasma membrane part	6.56E-05
GO:000577	vacuole	7.54E-05
GO:001602	integral to membrane	0.000111
GO:003122	intrinsic to membrane	0.000116
GO:004442	membrane part	0.000202
GO:002261	biological adhesion	0.000498

YY1

GO	GO as name	P-Value
GO:000563	nucleus	2.24E-05
GO:000613	nucleobase, nucleoside, nucleotide and	0.000684
GO:001602	membrane	-0.00242
GO:004442	membrane part	-0.00635
GO:000367	nucleic acid binding	0.00954

GO:000715	cell adhesion	0.000885
GO:000961	response to virus	0.00224

NFYA,B,C

GO	GO as name	P-Value
GO:000008	M phase of mitotic cell cycle	2.34E-26
GO:005130	cell division	2.34E-26
GO:000706	mitosis	4.37E-26
GO:000027	M phase	4.37E-26
GO:000704	cell cycle	7.70E-24
GO:000563	nucleus	5.87E-23
GO:002240	cell cycle phase	1.09E-22
GO:000027	mitotic cell cycle	1.97E-21
GO:004323	intracellular membrane-bound organelle	5.55E-20
GO:004322	membrane-bound organelle	5.71E-20

SNAI1-3

GO	GO as name	P-Value
GO:001602	membrane	7.40E-07
GO:000563	nucleus	-1.34E-06
GO:001046	gene expression	-1.77E-06
GO:004445	plasma membrane part	2.73E-06
GO:001602	integral to membrane	2.73E-06
GO:003122	intrinsic to membrane	3.19E-06
GO:000367	nucleic acid binding	-3.88E-06
GO:004442	membrane part	3.91E-06
GO:000588	integral to plasma membrane	5.90E-06
GO:003122	intrinsic to plasma membrane	8.52E-06

YY1

GO	GO as name	P-Value
GO:000563	nucleus	5.63E-11
GO:000613	nucleobase, nucleoside, nucleotide and	6.24E-09
GO:004323	intracellular membrane-bound organelle	7.66E-09
GO:004322	membrane-bound organelle	7.66E-09
GO:000367	nucleic acid binding	5.54E-07
GO:000625	DNA metabolic process	5.54E-07
GO:004322	intracellular organelle	5.79E-07
GO:004322	organelle	5.79E-07
GO:004423	primary metabolic process	5.79E-07

FOS,B,L1_JUNB,D

GO	GO as name	P-Value
GO:000758(	digestion	0.00054
GO:004425!	multicellular organismal macromolecule	0.00113
GO:004426ε	multicellular organismal protein metabo	0.00113
GO:003296:	collagen metabolic process	0.00113
GO:004425(	protein digestion	0.00113
GO:004425·	multicellular organismal protein cataboli	0.00113
GO:004426(	multicellular organismal macromolecule	0.00113
GO:003057·	collagen catabolic process	0.00113
GO:004424:	multicellular organismal catabolic proce:	0.00113
GO:004423(	multicellular organismal metabolic proce	0.00181

MYB

GO	GO as name	P-Value
GO:000706'	mitosis	3.15E-07
GO:000008'	M phase of mitotic cell cycle	3.15E-07
GO:000563·	nucleus	1.20E-06
GO:000027!	M phase	2.08E-06
GO:002240:	cell cycle phase	1.26E-05
GO:001563(	microtubule cytoskeleton	1.37E-05
GO:000027ε	mitotic cell cycle	1.44E-05
GO:002240:	cell cycle process	5.85E-05
GO:000704!	cell cycle	9.27E-05
GO:004443(	cytoskeletal part	0.000106

TBP

GO	GO as name	P-Value
GO:004873:	multicellular organismal development#s	1.31E-10
GO:003250:	multicellular organismal process	9.84E-10
GO:000715·	cell communication	2.38E-09
GO:004851:	multicellular organismal development#s	4.64E-08
GO:000727!	multicellular organismal development	1.88E-07
GO:004885(	anatomical structure development	3.08E-07
GO:000716!	signal transduction	4.39E-07
GO:000716(	cell surface receptor linked signal transd	2.93E-06
GO:003250:	developmental process	3.72E-05
GO:004442:	extracellular region part	5.37E-05

GO:004423'	cellular metabolic process	1.20E-06

FOS,B,L1_JUNB,D

GO	GO as name	P-Value
GO:004222:	response to chemical stimulus	1.40E-07
GO:004873:	multicellular organismal development#s	0.000122
GO:004851!	negative regulation of biological process	0.000151
GO:004852:	negative regulation of cellular process	0.000151
GO:004851:	multicellular organismal development#s	0.000182
GO:000716(	cell surface receptor linked signal transd	0.000849
GO:000960!	response to external stimulus	0.0011
GO:000761(	behavior	0.00114
GO:000485'	enzyme inhibitor activity	0.00114
GO:004885(	anatomical structure development	0.00123

MYB

GO	GO as name	P-Value
GO:000563·	nucleus	9.99E-15
GO:000027!	M phase	3.55E-11
GO:002240:	cell cycle phase	3.55E-11
GO:001563(	microtubule cytoskeleton	3.55E-11
GO:000706'	mitosis	8.07E-11
GO:000008'	M phase of mitotic cell cycle	1.16E-10
GO:000027ε	mitotic cell cycle	4.70E-10
GO:004443(	cytoskeletal part	7.27E-09
GO:004442ε	nuclear part	2.45E-08
GO:005130:	cell division	5.06E-08

TBP

GO	GO as name	P-Value
GO:004442:	extracellular region part	1.32E-28
GO:000960!	response to external stimulus	1.08E-14
GO:000961'	response to wounding	1.31E-12
GO:003250:	multicellular organismal process	1.33E-12
GO:000726'	cell-cell signaling	3.14E-12
GO:000510:	receptor binding	7.09E-12
GO:004873:	multicellular organismal development#s	1.11E-11
GO:000695:	immune response	1.32E-11
GO:000695:	defense response	1.54E-11
GO:000561!	extracellular space	6.86E-11

ZIC1-3

No GO term enrichment

ATF5_CREB3

GO	GO as name	P-Value
GO:000838	RNA splicing	0.000486
GO:001607	RNA metabolic process	0.000486
GO:000613	nucleobase, nucleoside, nucleotide and	0.000486
GO:000563	nucleus	0.000874
GO:000565	nuclear lumen#nucleoplasm	0.00101
GO:004328	biopolymer metabolic process	0.00111
GO:000639	RNA processing	0.00111
GO:000370	transcription factor activity	0.00111
GO:001046	gene expression	0.00238
GO:000639	mRNA processing	0.00349

PU.1

GO	GO as name	P-Value
GO:002261	biological adhesion	3.44E-14
GO:000715	cell adhesion	3.44E-14
GO:004322	membrane-bound organelle	-5.70E-14
GO:004323	intracellular membrane-bound organelle	-5.70E-14
GO:000588	plasma membrane	2.46E-11
GO:004322	organelle	-2.38E-09
GO:004322	intracellular organelle	-2.38E-09
GO:003122	intrinsic to plasma membrane	3.36E-09
GO:004445	plasma membrane part	1.07E-08
GO:000487	signal transducer activity	2.11E-08

RBPJ

No GO term enrichment

POU6F1

No GO term enrichment

SRF

No GO term enrichment

ZIC1-3

No GO term enrichment

ATF5_CREB3

GO	GO as name	P-Value
GO:000563	nucleus	2.35E-08
GO:000367	nucleic acid binding	3.66E-08
GO:000838	RNA splicing	3.66E-08
GO:000639	RNA processing	3.66E-08
GO:004442	nuclear part	1.85E-07
GO:000613	nucleobase, nucleoside, nucleotide and	1.33E-06
GO:000639	mRNA processing	2.91E-06
GO:001607	RNA metabolic process	2.91E-06
GO:001046	gene expression	5.21E-06
GO:004328	biopolymer metabolic process	9.80E-06

PU.1

GO	GO as name	P-Value
GO:002261	biological adhesion	6.45E-08
GO:000715	cell adhesion	6.45E-08
GO:001602	membrane	7.05E-07
GO:003122	intrinsic to membrane	3.28E-06
GO:001602	integral to membrane	3.28E-06
GO:004442	membrane part	1.46E-05
GO:004299	cell projection	1.46E-05
GO:000715	cell communication	2.28E-05
GO:003122	intrinsic to plasma membrane	2.64E-05
GO:000588	integral to plasma membrane	3.20E-05

RBPJ

No GO term enrichment

POU6F1

GO	GO as name	P-Value
GO:000588	plasma membrane	0.00665

SRF

GO	GO as name	P-Value
GO:000370	transcription factor activity	0.000407

TBX4,5

GO	GO as name	P-Value
GO:000370(	transcription factor activity	7.37e-05
GO:000635:	transcription, DNA-dependent	7.37e-05
GO:003277·	RNA biosynthetic process	7.37e-05
GO:000635(	transcription	7.37e-05
GO:004544!	regulation of transcription	0.000108
GO:000635!	regulation of transcription, DNA-depend	0.000148
GO:001921!	regulation of nucleobase, nucleoside, nu	0.000271
GO:001607(	RNA metabolic process	0.000469
GO:001046:	gene expression#regulation of gene expı	0.00058
GO:001922:	metabolic process#regulation of metabc	0.00186

EGR1-3

GO	GO as name	P-Value
GO:000996(	regulation of signal transduction	1.47E-05

EBF1

No GO term enrichment

FOXO1,3,4

No GO term enrichment

OCT4

GO	GO as name	P-Value
GO:000635!	regulation of transcription, DNA-depend	5.19e-08
GO:000563·	nucleus	7.41e-08
GO:000367'	DNA binding	7.41e-08
GO:000635:	transcription, DNA-dependent	7.41e-08
GO:003277·	RNA biosynthetic process	7.41e-08
GO:001921!	regulation of nucleobase, nucleoside, nu	7.41e-08
GO:004544!	regulation of transcription	1.74e-07
GO:000635(	transcription	5.77e-07
GO:000370(	transcription factor activity	6.26e-07
GO:005079·	cellular process#regulation of cellular pr	1.22e-06

TGIF1

TBX4,5

No GO term enrichment

EGR1-3

GO	GO as name	P-Value
GO:000996(	regulation of signal transduction	0.00909
GO:000813·	transcription factor binding	0.00909
GO:000726·	small GTPase mediated signal transducti	0.00909

EBF1

No GO term enrichment

FOXO1,3,4

No GO term enrichment

OCT4

GO	GO as name	P-Value
GO:000635!	regulation of transcription, DNA-depend	1.1e-08
GO:000635:	transcription, DNA-dependent	2.63e-08
GO:003277·	RNA biosynthetic process	2.63e-08
GO:004544!	regulation of transcription	2.63e-08
GO:001046:	gene expression#regulation of gene expı	9.9e-08
GO:001921!	regulation of nucleobase, nucleoside, nu	9.9e-08
GO:003132:	regulation of cellular metabolic process	2.8e-07
GO:000635(	transcription	2.8e-07
GO:001922:	metabolic process#regulation of metabc	1.37e-06
GO:004356!	sequence-specific DNA binding	1.71e-06

TGIF1

No GO term enrichment

NKX6-1,2
No GO term enrichment

BACH1,2

GO	GO as name	P-Value
GO:000588	plasma membrane	0.00145

GATA4

GO	GO as name	P-Value
GO:000563	nucleus	0.0043

No GO term enrichment

NKX6-1,2

GO	GO as name	P-Value
GO:003250	multicellular organismal process	2.42E-07
GO:000739	nervous system development	5.66E-05
GO:004873	multicellular organismal development#s	9.23E-05
GO:000727	multicellular organismal development	0.000141
GO:004885	anatomical structure development	0.00313
GO:000156	blood vessel development	0.00313
GO:000194	vasculature development	0.00313
GO:000715	cell communication	0.00314
GO:004851	blood vessel development#blood vessel	0.00601
GO:000176	generation of neurons#neuron migratio	0.00725

BACH1,2

GO	GO as name	P-Value
GO:000588	plasma membrane	2.84E-06
GO:000588	integral to plasma membrane	7.22E-05
GO:003122	intrinsic to plasma membrane	7.22E-05
GO:004851	multicellular organismal development#s	0.000174
GO:003250	multicellular organismal process	0.000396
GO:004222	response to chemical stimulus	0.000396
GO:004873	multicellular organismal development#s	0.000396
GO:004442	extracellular region part	0.00099
GO:004445	plasma membrane part	0.00177
GO:000761	behavior	0.00219

GATA4

GO	GO as name	P-Value
GO:000563	nucleus	1.04E-05
GO:000635	regulation of transcription, DNA-depend	1.04E-05
GO:000635	transcription, DNA-dependent	3.27E-05
GO:003277	RNA biosynthetic process	3.27E-05
GO:004544	regulation of transcription	3.27E-05
GO:001921	regulation of nucleobase, nucleoside, nu	4.49E-05
GO:000635	transcription	0.000122
GO:003132	regulation of cellular metabolic process	0.000157
GO:001607	RNA metabolic process	0.000234
GO:001046	gene expression#regulation of gene expr	0.000475

Supplementary Table 7 The sequences of siRNAs and RT-PCR primers (used to confirm knockdown efficiency)

					Percentage knockdown		
EntrezI	symbol	Sequence of siRNA	Forward primer sequence	Reverse primer sequence	1st	2nd	3rd
4602	MYB	GCCGCAGCCAUUCAGAGACACUAUA	GGCAGAAATCGCAAAGCTAC	ACCTTCCTGTTCGACCTTCC	81.7	93.3	93.4
1054	CEBPG	CAGAUGGCGACAAUGCAGGACAGUA	AACCATTGATCACCCTGCTC	AGTGCTGTTTTGCTGCGATA	91.8	80.2	93.5
3205	HOXA9	GCUUCCAGUCCAAGGCGACGGUGUU	AACAATGCTGAGAATGAGAGCGGC	TTTCCGAGTGGAGCGCGCATGAA	51.5	99.9	71.5
1050	CEBPA	CCUUCAACGACGAGUUCCUGGCCGA	AACCTTGTGCCTTGGAAATG	GAGGCAGGAAACCTCCAAAT	87.6	48.6	76.4
2313	FLI1	ACAUUAUGACCAAAGUGCACGGCAA	AGATCCGTATCAGATCCTGGGC	AGGAATTGCCACAGCTGGATCT	93.2	96.6	89.4
4300	MLLT3	CCUUAUAGAAGAAACUGGACACUUU	ATAGAGGAGGCAGCCGAAGT	TGGTGGAGGTTCGTGATGTA	87.3	90.7	91.2
2672	GFI1	UCUCCAGCCUCGGAGAAGUCAAUGU	AAGCAAGAAGGCTCACAGCTACCA	TGCATTTGAAGTGCTGTCTGCTCG	90.6	89.0	92.9
1869	E2F1	AGCCGUGGACUCUUCGGAGAACUUU	TGCTCTCCGAGGACACTGACA	GGGCTTTGATCACCATAACCATCTGC	83.0	85.0	85.0
2114	ETS2	GCCAACAGGCUUGGAUUCCAUUUCU	CAACAGGCTTGGATTCCATT	TTGACTCATCACAGCCTTGC	74.5	98.0	68.6
4851	NOTCH1	CCACCAGUUUGAAUGGUCAAUGCGA	TGAATGGCGGGAAGTGTGAAGC	GCACAGCTGCAGGCATAGTCT	90.1	87.9	91.8
2624	GATA2	CAGCAAGGCUCGUUCCUGUUCAGAA	AGCAAGGCTCGTTCCTGTT	CAGGCATTGCACAGGTAGTG	94.0	96.2	96.7
4779	NFE2L1	CCCAGCAAUUCUACCAGCCUCAACU	TGGAACAGCAGTGGCAAGATCTCA	GGCACTGTACAGGATTTCACTTGC	95.1	89.3	89.8
6720	SREBF1	UCAGAUACCACCAGCGUCUACCAUA	ATGGACGAGCCACCCTTC	CAAATAGGCCAGGGAAGTCA	85.6	83.9	70.9
333929	SNAI3	GGGCGUGUCUUCACCUGCAAGUACU	TCAAGATGCACATCCGCACTCACA	TTTGCAGATGGGCCCGAAGGTT	91.8	88.5	85.0
4605	MYBL2	CACCAGAAACGAGCCUGCCUUACAA	GAGGGATAGCAAGTGCAAGG	CAGGAACTTCCAGTCCTGCT	98.2	97.7	98.2
4609	MYC	CAGCGACUCUGAGGAGGAACAAGAA	AGCGACTCTGAGGAGGAACAAGAA	AGAAGGTGATCCAGACTCTGACCT	77.2	66.4	68.4
10472	ZNF238	AGACGUGCUAGCAGCUGCCAGUUAU	TTCATCTGAACAGCGACATTG	CGTCTTCAATGGGCAAGTCT	94.2	82.1	82.9
3394	IRF8	AGGUCUUCCGGAUGUUUCCAGAUAU	TTCGTACTGAGTTCGCTCCA	GGTTTATAGCCGCCAGTCAA	93.7	96.4	96.2
1958	EGR1	UCUCCCAGGACAAUUGAAAUUUGCU	CAGCAGCAGCACCTTCAAC	CTGGGGTAACTGGTCTCCAC	77.0	81.3	81.7
604	BCL6	GAGACCCAGUCUGAGUACUCAGAUU	GACTGTGAAGCAAGGCATTGGTGA	AACATCACTGGCATGGCGGGTGAAC	92.0	89.8	91.8
9935	MAFB	GCUACGCCCAGUCUUGCAGGUAUAA	CTGGCTTTCTGAACTTTGCGCGTT	TCCTTTCCTCGTTGCTCTCTTCCT	87.9	83.2	88.9
3665	IRF7	CCAAGGAGAAGAGCCUGGUCCUGGU	ATAACACCTGACCGCCACCTAACT	TGATCTCTCCAAGGAGCCACTCT	91.1	65.8	62.5
3397	ID1	CCUUCAGUUGGAGCUGAACUCGGAA	ACCCTCAACGGCGAGATCA	CTTCAGCGACACAAGATGCGAT	75.6	61.1	47.9
6615	SNAI1	AGGCCAAGGAUCUCCAGGCUCGAAA	TACAGCGAGCTGCAGGACTCTAAT	AGGACAGAGTCCCAGATGAGCATT	99.7	88.4	85.4
6688	PU.1	UAUAGAUCCGUGUCAUAGGGCACCA	GAAGACCTGGTGCCCTATGA	GGGGTGGAAGTCCCAGTAAT	98.2	97.7	98.4
6688	PU.1_2	AAUACUCGUGCGUUUGGCGUUGGUA	GAAGACCTGGTGCCCTATGA	GGGGTGGAAGTCCCAGTAAT	84.2	96.2	97.3
6667	SP1	GGAACAUCACCUUGCUACCUGUCAA	CAGTAGCAGCAGCACTGGAG	TTGCTGTTCTCATTGGGTGA	91.0	94.1	90.5
4297	MLL	AGUGGUUCCUGAGAAUGGAUUUGAA	CCAGTGATGATGGCTTTCAG	TTGATCGAGCTTCCTGGACT	70.2	56.9	66.6
4893	NRAS	GCGCACUGACAAUCCAGCUAAUCCA	AGCAAGTCATTTGCGGATATT	TCCTTGTTGGCAAATCACAC	90.2	91.8	95.0
7528	YY1	CCUUCGAUGGUUGUAAUAAGAAGUU	CAGAAGCAGGTGCAGATCAA	CAACCACTGTCTCATGGTCAA	86.9	93.7	89.6
4005	LMO2	GCAUUUCUGUGUAGGUGACAGAUAC	CTAGATCTGATGGGGGAAGC	AGCTACTGCAAGTTCAGGTTGA	95.0	97.0	97.3
4800	NFYA	GCCUGCUAUCCAAAGAAUCCCUCUA	TCTGATTGGGTTTCGGAGTC	TGGAGATCCTAGAAGGCTGTG	38.8	31.9	34.3
3209	HOXA13	GAUAUCAGCCACGACGAAUCUCUCU	AACGGCTGGAACGGCCAAATGTA	ATTGCACCTTGGTATAAGGCACGC	82.4	66.2	66.3
29128	UHRF1	GCCAGGUGGUCAUGCUCAACUACAA	GCCTGCAGAGGCTGTTCTAC	AGGAGCTGGATGGTGTCATT	92.3	94.8	95.7
10664	CTCF	CACACACAGGUACUCGUCCUCACAA	AGAACCAACCAGCCCAAACAGAAC	ATGTTCTCAATTGCACCTGTATTCTGGTCT	56.1	91.6	44.3
10155	TRIM28	GCCCUGAGACCAAACCUGUGCUUAU	AGGAGAAGTTGTCACCTCCCTACA	ACGTCTGCCTTGTCCTCAGTTA	98.3	95.2	95.0
2113	ETS1	GGAAUGUGCAGAUGUCCCACUAUUA	TGGACCAATCCAGCTATGGCAGTT	AGGCCACGGCTCAGTTTCTCATAA	87.8	88.4	90.1
6772	STAT1	CCUGUCACAGCUGGAUGAUCAAUAU	TGGAGCAGGTTCACCAGCTTTATG	TGAAACATCATTGGCAGCGTGCTC	96.1	81.7	82.9
2297	FOXD1	GAGCACUGAGAUGUCCGAUGCCUCU	TGACCCTGAGCACTGAGATG	CCTCTTCCTCGTCTTCTTCG	95.8	88.4	97.1
865	CBFB	UGAAUGGAGUCUGUGUUAUCUGGAA	ATTAAGTACACGGGCTTCAGG	GAGACAGATTGGTTCCTGTGG	95.6	96.1	96.4
9232	PTTG1	GCCUUAGAUGGGAGAUCUCAAGUUU	TGTGGTTGCTAAGGATGGGCTGAA	CTCTGTTGACAGTTCCCAAAGCCT	81.3	98.5	98.0

27086	FOXP1	GGCUGUGAAGCAGUGUGCGAAGAUU	CGATCCCTTCTCTGATTTGC	CATGCATAATGCCACAGGAC	88.4	82.2	84.5
3207	HOXA11	GCAGUCUCGUCCAAUUUCUAUAGCA	TCTTCCGGCCACACTGAGGACAA	AGACGCTGAAGAAGAACTCCCGTT	49.8	72.7	70.9
861	RUNX1	CACUAUCCAGGCGCCUUCACCUACU	TCAGGTTTGTCGGTCGAAGT	TGATGGCTCTGTGGTAGGTG	90.9	93.7	89.5
1052	CEBPD	ACAGCCUGGACUUACCACCACUAAA	ATCGACTTCAGCGCCTACAT	GCCTTGTGATTGCTGTTGAA	91.5	82.2	87.4
10732	TCFL5	AGUGGGAGAAGCAGCGCUAUGCAAA	TTGCCTGAGCAAGTTTGGAT	TTGTGTCCAACTGACGCATT	94.1	94.2	87.3
4790	NFKB1	CCAUCCUGGAACUACUAAAUCUAAU	ATGTATGTGAAGGCCCATCC	TGGTCCCACATAGTTGCAGA	98.8	98.4	98.4
648	BMI1	GGGUCAUCAGCAACUUCUUCUGGUU	CGACTTTTAACTTTCATTGTCTTTTC	CGTTGTTCGATGCATTTCTG	99.4	98.8	94.0
1051	CEBPB	CCUGAGUAAUCGCUUAAAGAUGUUC	CGTGTGTACACGGGACTGAC	CAACAAGCCCGTAGGAACAT	98.9	97.0	99.7
79191	IRX3	ACUGACGAGGAGGGAAACGCUUAUG	CTGGCCATCATCACCAAGAT	GCGCCCAAGTCATCTTATTC	95.1	93.5	97.4
4601	MXI1	GGAACGAAUACGAAUGGACAGCAUU	CCGGGCACAGAAACACAG	CGATTCTTTTCCAGCTCATTG	94.8	97.9	97.7
22887	FOXJ3	CCAGGUUCAGUUUGCCGAUCUUUGU	TAGAGAGGCTGGCAGTGGTT	CCAGGGTCATCCTTAGATCG	89.5	89.1	84.9
3206	HOXA10	CCGGGAGCUCACAGCCAACUUUAAU	CCCTGGGCAATTCCAAAGGTGAAA	AAACTCCTTCTCCAGCTCCAGTGT	82.8	78.0	82.8
2597	GAPDH	N/A	GAAGGTGAAGGTCGGAGTCA	AATGAAGGGGTAATTGATGG	N/A	N/A	N/A

Supplementary Table 8 All siRNAs tested and their differentiative effect.

		Down-regulated with siRNA			Up-regulated with siRNA			ALL changes					
Entrez ID	Gene symbol	Number of genes	%Also down in PMA	p-val	Number of genes	%Also up in PMA	p-val	Number of genes	%match response in PMA	p-val	Role in myeloid leukaemia	Fold change (0h vs 96h) PMA	Significant change Bstat>2.5, FC>2
4602	MYB	438	39.3%	1.68E-74	689	49.3%	2.86E-188	1127	45.4%	2.07E-121	Myeloid leukemia	0.11	Down
1054	CEBPG	130	23.1%	2.06E-07	103	35.0%	1.60E-13	233	28.3%	2.35E-05	-	0.40	-
4300	MLLT3	288	12.5%	1.09E-02	192	21.4%	7.79E-08	480	16.0%	2.32E-01	Myeloid leukemia	1.46	-
2313	FLI1	138	15.2%	5.23E-03	144	20.1%	1.64E-05	282	17.7%	4.70E-01	leukemia	0.49	Down
1050	CEBPA	167	24.6%	1.93E-10	249	19.3%	1.57E-07	416	21.4%	2.10E-02	Myeloid leukemia	0.58	-
3205	HOXA9	105	27.6%	4.91E-09	120	16.7%	3.60E-03	225	21.8%	5.49E-02	Myeloid leukemia	0.42	-
2672	GFI1	299	9.4%	2.83E-01	356	14.3%	3.76E-04	655	12.1%	6.90E-05	cancer	0.29	Down
3394	IRF8	204	15.2%	8.47E-04	107	11.2%	2.35E-01	311	13.8%	5.08E-02	-	0.07	Down
4609	MYC	314	13.4%	1.63E-03	136	8.8%	5.58E-01	450	12.0%	8.37E-04	Myeloid leukemia	0.35	Down
1869	E2F1	352	11.1%	4.59E-02	91	7.7%	4.27E-01	443	10.4%	1.74E-05	Myeloid leukemia	0.65	-
6688	SPI1(PU.1	1436	11.0%	2.21E-04	932	7.1%	1.97E-02	2368	9.5%	2.82E-34	Myeloid leukemia	1.63	Transient
1958	EGR1	537	8.2%	4.48E-01	198	4.5%	1.39E-02	735	7.2%	8.87E-17	Myeloid leukemia	0.95	Transient
2114	ETS2	503	12.5%	7.69E-04	715	4.2%	3.72E-07	1218	7.6%	2.22E-25	cancer	0.38	Transient
10472	ZNF238	253	10.3%	1.78E-01	363	4.1%	2.70E-04	616	6.7%	8.35E-16	-	0.93	Down
4800	NFYA	487	8.2%	4.61E-01	624	4.0%	7.76E-07	1111	5.9%	4.70E-33	-	0.56	-
3665	IRF7	618	9.7%	1.51E-01	798	3.9%	7.79E-09	1416	6.4%	4.15E-38	-	0.77	Transient
4779	NFE2L1	493	12.0%	3.63E-03	711	3.8%	3.08E-08	1204	7.1%	8.36E-28	-	2.36	Up
604	BCL6	795	10.6%	1.58E-02	676	3.7%	3.62E-08	1471	7.4%	1.91E-32	-	0.95	Up
4851	NOTCH1	477	11.1%	2.66E-02	532	3.6%	5.90E-07	1009	7.1%	2.96E-23	Myeloid leukemia	3.10	Up
9935	MAFB	824	9.6%	1.38E-01	169	3.6%	5.17E-03	993	8.6%	6.81E-17	-	67.73	Up
333929	SNAI3	396	10.1%	1.30E-01	648	3.4%	9.47E-09	1044	5.9%	1.73E-30	-	7.29	Up
6720	SREBF1	892	9.4%	1.56E-01	741	2.7%	1.57E-12	1633	6.4%	5.75E-45	cancer	0.51	Transient
29128	UHRF1	536	9.7%	1.71E-01	376	1.9%	7.40E-09	912	6.5%	5.33E-24	cancer	0.21	Down
6615	SNAI1	406	8.6%	4.47E-01	732	2.9%	1.07E-11	1138	4.9%	2.48E-40	cancer	2.26	Transient
7528	YY1	580	7.8%	2.96E-01	708	5.4%	1.71E-04	1288	6.4%	2.37E-34	cancer	0.54	-
3209	HOXA13	248	9.7%	2.82E-01	85	7.1%	3.55E-01	333	9.0%	7.56E-06	Myeloid leukemia	0.14	Down
6772	STAT1	353	8.8%	4.04E-01	679	3.7%	3.02E-08	1032	5.4%	3.41E-33	leukemia	0.70	Transient
2624	GATA2	172	10.5%	1.94E-01	100	3.0%	1.78E-02	272	7.7%	2.51E-06	leukemia	1.22	-
3397	ID1	84	13.1%	1.13E-01	104	6.7%	2.78E-01	188	9.6%	1.70E-03	leukemia	0.28	Down
6667	SP1	221	9.5%	3.00E-01	154	7.8%	3.73E-01	375	8.8%	1.06E-06	leukemia	0.56	-
4605	MYBL2	160	10.6%	1.85E-01	116	6.9%	2.81E-01	276	9.1%	5.12E-05	-	0.78	-

4297	MLL	178	9.6%	3.39E-01	70	7.1%	3.96E-01	248	8.9%	8.53E-05	Myeloid leukemia	0.56	-
10155	TRIM28	349	7.7%	3.53E-01	567	3.5%	1.89E-07	916	5.1%	4.79E-31	-	0.89	-
2113	ETS1	412	7.3%	2.18E-01	702	3.3%	8.64E-10	1114	4.8%	1.21E-40	cancer	3.08	Transient
27086	FOXP1	335	7.5%	2.91E-01	212	1.4%	3.58E-06	547	5.1%	9.94E-19	cancer	0.57	Down
4005	LMO2	157	8.9%	4.23E-01	473	3.2%	3.44E-07	630	4.6%	1.58E-23	leukemia	1.00	-
865	CBFB	250	7.6%	3.62E-01	203	3.0%	5.83E-04	453	5.5%	1.33E-14	Myeloid leukemia	0.51	-
10664	CTCF	247	7.7%	3.83E-01	105	8.6%	5.35E-01	352	8.0%	1.74E-07	cancer	0.43	-
4893	NRAS	361	6.9%	1.66E-01	566	7.1%	6.04E-02	927	7.0%	6.59E-22	Myeloid leukemia	1.07	-
4601	MXI1	443	2.9%	1.20E-06	431	1.2%	1.96E-12	874	2.1%	5.18E-51	cancer	1.08	Down
2297	FOXD1	201	7.0%	2.67E-01	244	5.3%	2.39E-02	445	6.1%	5.05E-13	-	0.43	Transient
861	RUNX1	356	6.2%	6.53E-02	399	2.8%	4.23E-07	755	4.4%	2.55E-29	Myeloid leukemia	0.42	Down
79191	IRX3	273	3.3%	4.45E-04	485	1.9%	3.81E-11	758	2.4%	9.68E-42	-	0.17	Down
1052	CEBPD	103	5.8%	2.21E-01	31	6.5%	4.68E-01	134	6.0%	7.87E-05	Myeloid leukemia	0.23	Down
648	BMI1	467	4.9%	1.90E-03	364	1.4%	5.63E-10	831	3.4%	1.61E-38	Myeloid leukemia	0.34	-
3206	HOXA10	143	3.5%	1.51E-02	76	1.3%	6.76E-03	219	2.7%	6.81E-12	leukemia	0.40	Down
1051	CEBPB	45	2.2%	9.60E-02	22	9.1%	6.87E-01	67	4.5%	1.50E-03	leukemia	4.21	Up
3207	HOXA11	88	5.7%	2.34E-01	57	3.5%	1.05E-01	145	4.8%	4.55E-06	leukemia	0.54	Transient
10732	TCFL5	184	4.9%	4.49E-02	133	1.5%	3.66E-04	317	3.5%	7.61E-15	-	0.27	Down
22887	FOXJ3	113	3.5%	3.23E-02	127	0.8%	8.73E-05	240	2.1%	2.26E-14	-	0.73	-
4790	NFKB1	362	4.1%	8.12E-04	606	2.3%	7.97E-12	968	3.0%	4.42E-48	leukemia	0.83	Transient
9232	PTTG1	185	4.3%	2.12E-02	610	3.8%	2.73E-07	795	3.9%	1.46E-33	cancer	0.41	Down
			9.4%	median		3.9%	median						
			6.4%	stdev		8.6%	stdev						
			15.8%	threshold		12.6%	threshold						

Highlighted fractions represent pro-differentiative changes greater than the median plus one std dev.

p-value was calculated using fisher's exact test on the numbers of genes perturbed above and the following numbers:
10824 genes are detected by Illumina during the timecourse
916 genes are downregulated with PMA (bstat>2.5, fold change>2)
967 genes are upregulated with PMA (bstat >2.5, fold change >2)

Note1: "myeloid leukemia". "leukemia" and "cancer" correponds to search terms found in entrez gene annotations
Note2: The last column, significant change refers to a reproducible significant difference between 0hr and 96hr (for up and down)
and between any 2 two other combination of timepoints (for transient). Significant change corresponds to a Bstat >2.5 and fold change>2)
no significant difference is flagged by (-)

Supplementary Table 9 Fourteen TFs that have differentiative overlap larger than 50%.

TFs	Overlap	P-value
MYB	0.69	<0.001
CEBPG	0.63	<0.001
E2F1	0.58	0.004
MLL	0.55	0.026
HOXA9	0.55	0.055
EGR1	0.55	0.064
CEBPA	0.55	0.081
GATA2	0.55	0.076
MYC	0.52	0.242
FLI1	0.52	0.254
MLLT3	0.52	0.245
YY1	0.51	0.34
IRF8	0.51	0.377
GFI1	0.51	0.42
NCs4	0.5	0.5
NCs3	0.5	0.5
NCs2	0.5	0.5
NC 0	0.5	0.5

Also shown are the differentiative overlap with 4 different negative control samples (NC 0 and NCs2, 3, and 4). Third column shows the p-value for the overlap under a permutation test.

Supplementary Table 10 Confirmation of surface marker changes induced by siRNA or PMA by flow cytometry.
Average given in table represents the average of MFI (mean fluorescence intensity) of 3 identical wells.

	siRNA effect							PMA effect			
Antibody	NC siRNA Average	MYB siRNA	Fold chang	p value t-test	GFI1 siRNA	Fold chang	p value t-test	untreated average	PMA average	Fold chang	p value t-test
CD9-F	295	239	0.81	0.1627	319	1.08	0.4307	334	436	1.31	0.0355
CD15-F	344	263	0.76	0.1795	399	1.16	0.0293	484	314	0.65	0.1542
CD18-F	545	596	1.09	0.3477	627	1.15	0.1184	655	1623	2.48	0.0105
HLA-DR-F	1346	1188	0.88	0.5214	1474	1.10	0.0102	1009	523	0.52	0.1985
MPO-F	1033	1030	1.00	0.8816	1127	1.09	0.0483	1091	387	0.35	0.2263
CD105-F	2031	2289	1.13	0.0372	2185	1.08	0.1580	1880	1202	0.64	0.2105
CD11b-PE	682	933	1.37	0.1221	842	1.23	0.0019	771	3340	4.33	0.0716
CD11c-PE	368	349	0.95	0.7184	335	0.91	0.1765	247	2062	8.37	0.0349
CD54-PE	169	207	1.22	0.0217	174	1.03	0.1898	194	5160	26.66	0.0182
CD14-PE	321	890	2.77	0.0136	368	1.15	0.1152	450	2223	4.94	0.0414
CCR2-Alexa647	823	761	0.92	0.4677	875	1.06	0.3183	982	-23	-0.02	0.0225

Note significant upregulation of CD11b (ITGAM), CD54 and CD14 with MYB siRNA. In comparison GFI1 siRNA induces CD11b but not CD54 or CD14.

Supplementary Table 11 Relationship between pro-differentiative changes induced by MYB siRNA and other siRNAs.

	CEBPA	CEBPG	FLI1	GFI1	HOXA9	MLLT3	MYB
PMA-like changes shared with MYB	53%	88%	47%	65%	66%	64%	100%
predicted target of MYB	no	yes	yes	no	no	no	yes
Fold change upon MYB KD	1.17	0.74	1.17	0.62	0.85	1.02	0.68
b-stat with MYB siRNA	-5.38	0.01	-5.61	20.80	-4.29	-7.35	20.27

Note: MYB KD induces a significant decrease in CEBPG and GFI1 levels (B-statistic >0)

Nature Genetics: doi:10.1038/ng.375

Supplementary Table 12 Transcription factors detected in THP-1 cells during differentiation

Symbol	EntrezID	CAGE detected	illumina detected	expression
AFF3	3899	yes	yes	dynamic
AKNA	80709	yes	yes	dynamic
AR	367	yes	yes	dynamic
ARID2	196528	yes	yes	dynamic
ARID3A	1820	yes	yes	dynamic
ARID5B	84159	yes	yes	dynamic
ATF3	467	yes	yes	dynamic
ATF5	22809	yes	yes	dynamic
BAPX1	579	yes	yes	dynamic
BATF	10538	yes	yes	dynamic
BCL6	604	yes	yes	dynamic
BCOR	54880	yes	yes	dynamic
BHLHB2	8553	yes	yes	dynamic
BHLHB3	79365	yes	yes	dynamic
CBFA2T3	863	yes	yes	dynamic
CDCA7L	55536	yes	yes	dynamic
CEBPB	1051	yes	yes	dynamic
CEBPD	1052	yes	yes	dynamic
CITED2	10370	yes	yes	dynamic
CNOT10	25904	yes	yes	dynamic
CREB3L4	148327	yes	yes	dynamic
DATF1	11083	yes	yes	dynamic
DBP	1628	yes	yes	dynamic
DDIT3	1649	yes	yes	dynamic
DEDD2	162989	yes	yes	dynamic
DEK	7913	yes	yes	dynamic
DLX1	1745	yes	yes	dynamic
DLX3	1747	yes	yes	dynamic
DMRT2	10655	yes	yes	dynamic
DMRTA2	63950	yes	yes	dynamic
E2F2	1870	yes	yes	dynamic
E2F6	1876	yes	yes	dynamic
EGR1	1958	yes	yes	dynamic
EGR2	1959	yes	yes	dynamic
EGR3	1960	yes	yes	dynamic

Readme:
Detection in CAGE is defined as the average signal for the three biological replicates was greater than 3 tags per million (TPM)
Detection in Illumina microarray required the average detection score of the three biological replciates to be less than 0.01.

Dynamic expression - required B-statistic >2.5 and fold change >2 using illumina arrays.
Contrasts tested were undifferentiated vs each time point, and time-point vs consecutive timepoint.

Curated transcription factor list reference:
3. Roach, J.C. et al. Transcription factor expression in lipopolysaccharide-activated peripheral-blood-derived mononuclear cells. Proc Natl Acad Sci U S A 104, 16245-50 (2007).

EGR4	1961	yes	yes	dynamic
ETS1	2113	yes	yes	dynamic
ETS2	2114	yes	yes	dynamic
ETV3	2117	yes	yes	dynamic
ETV5	2119	yes	yes	dynamic
EZH2	2146	yes	yes	dynamic
FLI1	2313	yes	yes	dynamic
FOS	2353	yes	yes	dynamic
FOSB	2354	yes	yes	dynamic
FOSL1	8061	yes	yes	dynamic
FOXD1	2297	yes	yes	dynamic
FOXM1	2305	yes	yes	dynamic
FOXO1A	2308	yes	yes	dynamic
FOXP1	27086	yes	yes	dynamic
FUBP1	8880	yes	yes	dynamic
GLI4	2738	yes	yes	dynamic
HDAC1	3065	yes	yes	dynamic
HDAC4	9759	yes	yes	dynamic
HES1	3280	yes	yes	dynamic
HEYL	26508	yes	yes	dynamic
HIVEP1	3096	yes	yes	dynamic
HIVEP2	3097	yes	yes	dynamic
HLX1	3142	yes	yes	dynamic
HLXB9	3110	yes	yes	dynamic
HMGA1	3159	yes	yes	dynamic
HMGB2	3148	yes	yes	dynamic
HMGN1	3150	yes	yes	dynamic
HOP	84525	yes	yes	dynamic
HOXA10	3206	yes	yes	dynamic
HOXA11	3207	yes	yes	dynamic
HOXA13	3209	yes	yes	dynamic
HTATSF1	27336	yes	yes	dynamic
ID1	3397	yes	yes	dynamic
ID2	3398	yes	yes	dynamic
IFI16	3428	yes	yes	dynamic
IRF2BP2	359948	yes	yes	dynamic
IRF7	3665	yes	yes	dynamic
IRF8	3394	yes	yes	dynamic

IRX3	79191	yes	yes	dynamic
ISGF3G	10379	yes	yes	dynamic
JARID1B	10765	yes	yes	dynamic
KIAA0284	283638	yes	yes	dynamic
KIAA1443	57594	yes	yes	dynamic
KLF10	7071	yes	yes	dynamic
KLF11	8462	yes	yes	dynamic
KLF2	10365	yes	yes	dynamic
KLF9	687	yes	yes	dynamic
LASS4	79603	yes	yes	dynamic
LMO4	8543	yes	yes	dynamic
LOC401074	401074	yes	yes	dynamic
MAFB	9935	yes	yes	dynamic
MAFF	23764	yes	yes	dynamic
MEF2D	4209	yes	yes	dynamic
MITF	4286	yes	yes	dynamic
MSC	9242	yes	yes	dynamic
MXD1	4084	yes	yes	dynamic
MXD3	83463	yes	yes	dynamic
MXI1	4601	yes	yes	dynamic
MYB	4602	yes	yes	dynamic
MYC	4609	yes	yes	dynamic
MYCBP	26292	yes	yes	dynamic
NAB2	4665	yes	yes	dynamic
NCOA3	8202	yes	yes	dynamic
NCOA7	135112	yes	yes	dynamic
NFATC1	4772	yes	yes	dynamic
NFE2	4778	yes	yes	dynamic
NFE2L1	4779	yes	yes	dynamic
NFIX	4784	yes	yes	dynamic
NFKB1	4790	yes	yes	dynamic
NFKB2	4791	yes	yes	dynamic
NFYB	4801	yes	yes	dynamic
NR1H3	10062	yes	yes	dynamic
NR2F6	2063	yes	yes	dynamic
NR3C1	2908	yes	yes	dynamic
NR4A1	3164	yes	yes	dynamic
NRIP3	56675	yes	yes	dynamic

ONECUT2	9480	yes	yes	dynamic
PAXIP1	22976	yes	yes	dynamic
PCBD1	5092	yes	yes	dynamic
PHC2	1912	yes	yes	dynamic
PHC3	80012	yes	yes	dynamic
PHF10	55274	yes	yes	dynamic
PHF13	148479	yes	yes	dynamic
PHF15	23338	yes	yes	dynamic
PHF16	9767	yes	yes	dynamic
PHF17	79960	yes	yes	dynamic
PLAGL2	5326	yes	yes	dynamic
PML	5371	yes	yes	dynamic
POU2F1	5451	yes	yes	dynamic
POU2F2	5452	yes	yes	dynamic
PPARD	5467	yes	yes	dynamic
PPARG	5468	yes	yes	dynamic
PRDM1	639	yes	yes	dynamic
PSIP1	11168	yes	yes	dynamic
RARA	5914	yes	yes	dynamic
RELB	5971	yes	yes	dynamic
REPIN1	29803	yes	yes	dynamic
RERE	473	yes	yes	dynamic
RFP2	10206	yes	yes	dynamic
RFX5	5993	yes	yes	dynamic
RNF24	11237	yes	yes	dynamic
RUNX1	861	yes	yes	dynamic
RYBP	23429	yes	yes	dynamic
SCMH1	22955	yes	yes	dynamic
SERTAD1	29950	yes	yes	dynamic
SERTAD2	9792	yes	yes	dynamic
SFMBT1	51460	yes	yes	dynamic
SIN3A	25942	yes	yes	dynamic
SIX5	147912	yes	yes	dynamic
SLC2A4RG	56731	yes	yes	dynamic
SMAD3	4088	yes	yes	dynamic
SMARCA2	6595	yes	yes	dynamic
SMARCAL1	50485	yes	yes	dynamic
SNAI1	6615	yes	yes	dynamic

SNAI3	333929	yes	yes	dynamic
SNFT	55509	yes	yes	dynamic
SOX12	6666	yes	yes	dynamic
SOX4	6659	yes	yes	dynamic
SP4	6671	yes	yes	dynamic
SPI1	6688	yes	yes	dynamic
SPOCD1	90853	yes	yes	dynamic
SREBF1	6720	yes	yes	dynamic
SREBF2	6721	yes	yes	dynamic
STAT1	6772	yes	yes	dynamic
STAT2	6773	yes	yes	dynamic
TAF1B	9014	yes	yes	dynamic
TARDBP	23435	yes	yes	dynamic
TCF19	6941	yes	yes	dynamic
TCF4	6925	yes	yes	dynamic
TCFL5	10732	yes	yes	dynamic
TEF	7008	yes	yes	dynamic
TFAP4	7023	yes	yes	dynamic
TFB1M	51106	yes	yes	dynamic
TFDP1	7027	yes	yes	dynamic
TFE3	7030	yes	yes	dynamic
TFEB	7942	yes	yes	dynamic
TFPT	29844	yes	yes	dynamic
TGIF	7050	yes	yes	dynamic
TGIF2	60436	yes	yes	dynamic
THOC4	10189	yes	yes	dynamic
THRA	7067	yes	yes	dynamic
TLE1	7088	yes	yes	dynamic
TLE3	7090	yes	yes	dynamic
TRIM25	7706	yes	yes	dynamic
TRIM32	22954	yes	yes	dynamic
TWIST1	7291	yes	yes	dynamic
UHRF1	29128	yes	yes	dynamic
USF2	7392	yes	yes	dynamic
VDR	7421	yes	yes	dynamic
WBSCR14	51085	yes	yes	dynamic
WT1	7490	yes	yes	dynamic
XBP1	7494	yes	yes	dynamic

YEATS4	8089	yes	yes	dynamic
ZFP36L1	677	yes	yes	dynamic
ZFP95	23660	yes	yes	dynamic
ZHX2	22882	yes	yes	dynamic
ZHX3	23051	yes	yes	dynamic
ZNF174	7727	yes	yes	dynamic
ZNF256	10172	yes	yes	dynamic
ZNF274	10782	yes	yes	dynamic
ZNF278	23598	yes	yes	dynamic
ZNF281	23528	yes	yes	dynamic
ZNF297B	23099	yes	yes	dynamic
ZNF473	25888	yes	yes	dynamic
AATF	26574	yes	yes	static
ADNP	23394	yes	yes	static
AFF1	4299	yes	yes	static
AFF4	27125	yes	yes	static
AHCTF1	25909	yes	yes	static
AHR	196	yes	yes	static
AIP	9049	yes	yes	static
ARID1A	8289	yes	yes	static
ARID4A	5926	yes	yes	static
ARID4B	51742	yes	yes	static
ARID5A	10865	yes	yes	static
ARNT	405	yes	yes	static
ARNTL	406	yes	yes	static
ASXL1	171023	yes	yes	static
ASXL2	55252	yes	yes	static
ATF2	1386	yes	yes	static
ATF4	468	yes	yes	static
ATF6	22926	yes	yes	static
BARD1	580	yes	yes	static
BARX1	56033	yes	yes	static
BAZ1A	11177	yes	yes	static
BAZ1B	9031	yes	yes	static
BCLAF1	9774	yes	yes	static
BDP1	55814	yes	yes	static
BRD1	23774	yes	yes	static
BRD4	23476	yes	yes	static

BRD8	10902	yes	yes	static
BRD9	65980	yes	yes	static
BRF1	2972	yes	yes	static
BRF2	55290	yes	yes	static
BRPF3	27154	yes	yes	static
BTAF1	9044	yes	yes	static
BTF3L4	91408	yes	yes	static
CBFA2T2	9139	yes	yes	static
CBFB	865	yes	yes	static
CCNH	902	yes	yes	static
CCNT1	904	yes	yes	static
CCNT2	905	yes	yes	static
CDK7	1022	yes	yes	static
CDK9	1025	yes	yes	static
CEBPA	1050	yes	yes	static
CEBPG	1054	yes	yes	static
CEBPZ	10153	yes	yes	static
CHES1	1112	yes	yes	static
CHFR	55743	yes	yes	static
CNOT1	23019	yes	yes	static
CNOT2	4848	yes	yes	static
CNOT3	4849	yes	yes	static
CNOT4	4850	yes	yes	static
CNOT7	29883	yes	yes	static
CNOT8	9337	yes	yes	static
COBRA1	25920	yes	yes	static
COPS5	10987	yes	yes	static
CREB1	1385	yes	yes	static
CREB3	10488	yes	yes	static
CREB3L2	64764	yes	yes	static
CREBBP	1387	yes	yes	static
CREBL1	1388	yes	yes	static
CREG1	8804	yes	yes	static
CRI1	23741	yes	yes	static
CRSP2	9282	yes	yes	static
CRSP3	9439	yes	yes	static
CRSP7	9441	yes	yes	static
CRSP8-ambiguous	9442	yes	yes	static

CRSP9	9443	yes	yes	static
CSDA	8531	yes	yes	static
CTBP1	1487	yes	yes	static
CTCF	10664	yes	yes	static
CTNNB1	1499	yes	yes	static
CUTL1	1523	yes	yes	static
DAXX	1616	yes	yes	static
DCP1A	55802	yes	yes	static
DEDD	9191	yes	yes	static
DMTF1	9988	yes	yes	static
DPF2	5977	yes	yes	static
DR1	1810	yes	yes	static
DRAP1	10589	yes	yes	static
E2F1	1869	yes	yes	static
E2F3	1871	yes	yes	static
E2F4	1874	yes	yes	static
E2F7	144455	yes	yes	static
E4F1	1877	yes	yes	static
EBF3	253738	yes	yes	static
EDF1	8721	yes	yes	static
EED	8726	yes	yes	static
ELF1	1997	yes	yes	static
ELF2	1998	yes	yes	static
ELF4	2000	yes	yes	static
ELK1	2002	yes	yes	static
EP300	2033	yes	yes	static
EPAS1	2034	yes	yes	static
EPC1	80314	yes	yes	static
ERF	2077	yes	yes	static
ESRRA	2101	yes	yes	static
ETV6	2120	yes	yes	static
EZH1	2145	yes	yes	static
FALZ	2186	yes	yes	static
FAM48A	55578	yes	yes	static
FLJ21616	79618	yes	yes	static
FOSL2	2355	yes	yes	static
FOXJ2	55810	yes	yes	static
FOXO3A	2309	yes	yes	static

FOXP4	116113	yes	yes	static
FUBP3	8939	yes	yes	static
GABPA	2551	yes	yes	static
GABPB2	2553	yes	yes	static
GATA2	2624	yes	yes	static
GATAD2B	57459	yes	yes	static
GCN5L2	2648	yes	yes	static
GMEB2	26205	yes	yes	static
GTF2A1	2957	yes	yes	static
GTF2A2	2958	yes	yes	static
GTF2B	2959	yes	yes	static
GTF2E1	2960	yes	yes	static
GTF2E2	2961	yes	yes	static
GTF2F1	2962	yes	yes	static
GTF2F2	2963	yes	yes	static
GTF2H1	2965	yes	yes	static
GTF2H3	2967	yes	yes	static
GTF2I	2969	yes	yes	static
GTF2IRD1	9569	yes	yes	static
GTF3A	2971	yes	yes	static
GTF3C1	2975	yes	yes	static
GTF3C2	2976	yes	yes	static
GTF3C3	9330	yes	yes	static
GTF3C5	9328	yes	yes	static
HBP1	26959	yes	yes	static
HCFC1	3054	yes	yes	static
HCFC2	29915	yes	yes	static
HDAC2	3066	yes	yes	static
HDAC3	8841	yes	yes	static
HDAC5	10014	yes	yes	static
HDAC7A	51564	yes	yes	static
HDAC8	55869	yes	yes	static
HDAC9	9734	yes	yes	static
HES6	55502	yes	yes	static
HHEX	3087	yes	yes	static
HIF1A	3091	yes	yes	static
HKR1	284459	yes	yes	static
HKR3	3104	yes	yes	static

HMG20A	10363	yes	yes	static
HMG20B	10362	yes	yes	static
HMGN2	3151	yes	yes	static
HMGN3	9324	yes	yes	static
HNF4G	3174	yes	yes	static
HOXA9	3205	yes	yes	static
HSBP1	3281	yes	yes	static
HSF1	3297	yes	yes	static
HSF2	3298	yes	yes	static
HTLF	3344	yes	yes	static
ILF3	3609	yes	yes	static
IRF1	3659	yes	yes	static
IRF2	3660	yes	yes	static
IRF2BP1	26145	yes	yes	static
IRF3	3661	yes	yes	static
IRF5	3663	yes	yes	static
JARID1A	5927	yes	yes	static
JUN	3725	yes	yes	static
JUND	3727	yes	yes	static
KLF13	51621	yes	yes	static
KLF16	83855	yes	yes	static
KLF4	9314	yes	yes	static
KLF6	1316	yes	yes	static
LASS5	91012	yes	yes	static
LASS6	253782	yes	yes	static
LBX2	85474	yes	yes	static
LMO2	4005	yes	yes	static
LRRFIP1	9208	yes	yes	static
LYL1	4066	yes	yes	static
LZTR1	8216	yes	yes	static
MAFG	4097	yes	yes	static
MAML3	55534	yes	yes	static
MAX	4149	yes	yes	static
MBD1	4152	yes	yes	static
MBD2	8932	yes	yes	static
MECP2	4204	yes	yes	static
MED4	29079	yes	yes	static
MEF2A	4205	yes	yes	static

MEF2B	4207	yes	yes	static
MEF2C	4208	yes	yes	static
MEIS2	4212	yes	yes	static
MIZF	25988	yes	yes	static
MKRN1	23608	yes	yes	static
MKRN2	23609	yes	yes	static
MLL	4297	yes	yes	static
MLLT6	4302	yes	yes	static
MLLT7	4303	yes	yes	static
MLR2	84458	yes	yes	static
MLX	6945	yes	yes	static
MNAT1	4331	yes	yes	static
MTA1	9112	yes	yes	static
MTA2	9219	yes	yes	static
MTA3	57504	yes	yes	static
MTF1	4520	yes	yes	static
MTF2	22823	yes	yes	static
MXD4	10608	yes	yes	static
MYBBP1A	10514	yes	yes	static
MYBL2	4605	yes	yes	static
MYCL1	4610	yes	yes	static
MYEF2	50804	yes	yes	static
MYNN	55892	yes	yes	static
NBL1	4681	yes	yes	static
NCOA4	8031	yes	yes	static
NCOA5	57727	yes	yes	static
NCOR1	9611	yes	yes	static
NCOR2	9612	yes	yes	static
NFAT5	10725	yes	yes	static
NFATC3	4775	yes	yes	static
NFE2L2	4780	yes	yes	static
NFIC	4782	yes	yes	static
NFRKB	4798	yes	yes	static
NFX1	4799	yes	yes	static
NFYA	4800	yes	yes	static
NFYC	4802	yes	yes	static
NKRF	55922	yes	yes	static
NMI	9111	yes	yes	static

NPAT	4863	yes	yes	static
NR1D2	9975	yes	yes	static
NR1H2	7376	yes	yes	static
NR2C1	7181	yes	yes	static
NR2C2	7182	yes	yes	static
NR4A2	4929	yes	yes	static
NR4A3	8013	yes	yes	static
NRBF2	29982	yes	yes	static
NSD1	64324	yes	yes	static
OLIG2	10215	yes	yes	static
OTX1	5013	yes	yes	static
PBX2	5089	yes	yes	static
PBX3	5090	yes	yes	static
PBXIP1	57326	yes	yes	static
PCGF1	84759	yes	yes	static
PCGF4	648	yes	yes	static
PCGF6	84108	yes	yes	static
PCQAP	51586	yes	yes	static
PER2	8864	yes	yes	static
PHF1	5252	yes	yes	static
PHF11	51131	yes	yes	static
PHF12	57649	yes	yes	static
PHF14	9678	yes	yes	static
PHF19	26147	yes	yes	static
PHF2	5253	yes	yes	static
PHF3	23469	yes	yes	static
PHF7	51533	yes	yes	static
PKNOX1	5316	yes	yes	static
PLAGL1	5325	yes	yes	static
POLR2A	5430	yes	yes	static
POLR3A	11128	yes	yes	static
POLR3E	55718	yes	yes	static
POLR3H	171568	yes	yes	static
PPARBP	5469	yes	yes	static
PPARGC1B	133522	yes	yes	static
PQBP1	10084	yes	yes	static
PRDM10	56980	yes	yes	static
PRDM4	11108	yes	yes	static

PREB	10113	yes	yes	static
PRKCBP1	23613	yes	yes	static
PURA	5813	yes	yes	static
RB1	5925	yes	yes	static
RBBP4	5928	yes	yes	static
RBL1	5933	yes	yes	static
RBL2	5934	yes	yes	static
RBPSUH	3516	yes	yes	static
RDBP	7936	yes	yes	static
REL	5966	yes	yes	static
RELA	5970	yes	yes	static
REXO1	57455	yes	yes	static
RFP	5987	yes	yes	static
RFX1	5989	yes	yes	static
RFX2	5990	yes	yes	static
RFX3	5991	yes	yes	static
RFXANK	8625	yes	yes	static
RING1	6015	yes	yes	static
RNF4	6047	yes	yes	static
RQCD1	9125	yes	yes	static
RUNX2	860	yes	yes	static
RUNX3	864	yes	yes	static
RXRA	6256	yes	yes	static
RXRB	6257	yes	yes	static
SAFB	6294	yes	yes	static
SAFB2	9667	yes	yes	static
SATB2	23314	yes	yes	static
SFMBT2	57713	yes	yes	static
SHOX2	6474	yes	yes	static
SIAHBP1	22827	yes	yes	static
SIN3B	23309	yes	yes	static
SMAD2	4087	yes	yes	static
SMAD4	4089	yes	yes	static
SMARCA3	6596	yes	yes	static
SMARCA4	6597	yes	yes	static
SMARCA5	8467	yes	yes	static
SMARCB1	6598	yes	yes	static
SMARCC1	6599	yes	yes	static

SMARCC2	6601	yes	yes	static
SMARCD1	6602	yes	yes	static
SMARCD2	6603	yes	yes	static
SMARCD3	6604	yes	yes	static
SMARCE1	6605	yes	yes	static
SNAPC1	6617	yes	yes	static
SNAPC3	6619	yes	yes	static
SND1	27044	yes	yes	static
SNW1	22938	yes	yes	static
SOLH	6650	yes	yes	static
SON	6651	yes	yes	static
SOX13	9580	yes	yes	static
SOX30	11063	yes	yes	static
SP1	6667	yes	yes	static
SP2	6668	yes	yes	static
SP3	6670	yes	yes	static
SRCAP	10847	yes	yes	static
SRF	6722	yes	yes	static
SSRP1	6749	yes	yes	static
STAT3	6774	yes	yes	static
STAT5A	6776	yes	yes	static
STAT5B	6777	yes	yes	static
STAT6	6778	yes	yes	static
SUB1	10923	yes	yes	static
SUPT16H	11198	yes	yes	static
SUPT4H1	6827	yes	yes	static
SUPT5H	6829	yes	yes	static
SUPT6H	6830	yes	yes	static
SURB7	9412	yes	yes	static
TADA3L	10474	yes	yes	static
TAF1	6872	yes	yes	static
TAF10	6881	yes	yes	static
TAF12	6883	yes	yes	static
TAF13	6884	yes	yes	static
TAF15	8148	yes	yes	static
TAF1A	9015	yes	yes	static
TAF1C	9013	yes	yes	static
TAF2	6873	yes	yes	static

TAF4	6874	yes	yes	static
TAF5	6877	yes	yes	static
TAF5L	27097	yes	yes	static
TAF6	6878	yes	yes	static
TAF7	6879	yes	yes	static
TAF9	6880	yes	yes	static
TAF9L	51616	yes	yes	static
TBP	6908	yes	yes	static
TBPL1	9519	yes	yes	static
TBX18	9096	yes	yes	static
TCEA1	6917	yes	yes	static
TCEA2	6919	yes	yes	static
TCEB1	6921	yes	yes	static
TCEB2	6923	yes	yes	static
TCERG1	10915	yes	yes	static
TCF12	6938	yes	yes	static
TFAM	7019	yes	yes	static
TFAP2C	7022	yes	yes	static
TFB2M	64216	yes	yes	static
TFCP2	7024	yes	yes	static
TGFB1I1	7041	yes	yes	static
TH1L	51497	yes	yes	static
THRAP2	23389	yes	yes	static
THRAP3	9967	yes	yes	static
THRAP4	9862	yes	yes	static
TLE4	7091	yes	yes	static
TMF1	7110	yes	yes	static
TOPORS	10210	yes	yes	static
TRIM24	8805	yes	yes	static
TRIM28	10155	yes	yes	static
TRPS1	7227	yes	yes	static
TRRAP	8295	yes	yes	static
UBN1	29855	yes	yes	static
UBP1	7342	yes	yes	static
UBTF	7343	yes	yes	static
VAX2	25806	yes	yes	static
VPS72	6944	yes	yes	static
WHSC2	7469	yes	yes	static

XAB1	11321	yes	yes	static
YBX1	4904	yes	yes	static
YY1	7528	yes	yes	static
ZBTB25	7597	yes	yes	static
ZBTB7B	51043	yes	yes	static
ZF	58487	yes	yes	static
ZFHX1B	9839	yes	yes	static
ZFP161	7541	yes	yes	static
ZHX1	11244	yes	yes	static
ZNF133	7692	yes	yes	static
ZNF134	7693	yes	yes	static
ZNF142	7701	yes	yes	static
ZNF143	7702	yes	yes	static
ZNF148	7707	yes	yes	static
ZNF16	7564	yes	yes	static
ZNF160	90338	yes	yes	static
ZNF161	7716	yes	yes	static
ZNF202	7753	yes	yes	static
ZNF207	7756	yes	yes	static
ZNF211	10520	yes	yes	static
ZNF213	7760	yes	yes	static
ZNF217	7764	yes	yes	static
ZNF219	51222	yes	yes	static
ZNF224	7767	yes	yes	static
ZNF24	7572	yes	yes	static
ZNF263	10127	yes	yes	static
ZNF268	10795	yes	yes	static
ZNF277	11179	yes	yes	static
ZNF295	49854	yes	yes	static
ZNF297	9278	yes	yes	static
ZNF336	64412	yes	yes	static
ZNF35	7584	yes	yes	static
ZNF38	7589	yes	yes	static
ZNF384	171017	yes	yes	static
ZNF398	57541	yes	yes	static
ZNF42	7593	yes	yes	static
ZNF444	55311	yes	yes	static
ZNF524	147807	yes	yes	static

ZNF655	79027	yes	yes	static
ZNF668	79759	yes	yes	static
ZNF91	7644	yes	yes	static
ZNFN1A1	10320	yes	yes	static
ZNRD1	30834	yes	yes	static
ARNT2	9915	no	yes	dynamic
BIN1	274	no	yes	dynamic
CART1	8092	no	yes	dynamic
CUTL2	23316	no	yes	dynamic
DMRT3	58524	no	yes	dynamic
EOMES	8320	no	yes	dynamic
FOXF2	2295	no	yes	dynamic
GTF2IRD2P	401375	no	yes	dynamic
HES4	57801	no	yes	dynamic
HESX1	8820	no	yes	dynamic
HEY2	23493	no	yes	dynamic
HIC2	23119	no	yes	dynamic
HOXA5	3202	no	yes	dynamic
HOXB4	3214	no	yes	dynamic
HSF2BP	11077	no	yes	dynamic
IRX5	10265	no	yes	dynamic
LHX2	9355	no	yes	dynamic
LZTFL1	54585	no	yes	dynamic
MESP1	55897	no	yes	dynamic
MNT	4335	no	yes	dynamic
NFIL3	4783	no	yes	dynamic
NKX2-2	4821	no	yes	dynamic
OLIG1	116448	no	yes	dynamic
POU4F1	5457	no	yes	dynamic
RFXAP	5994	no	yes	dynamic
SCML2	10389	no	yes	dynamic
SOX18	54345	no	yes	dynamic
TCF3	6929	no	yes	dynamic
TFAP2A	7020	no	yes	dynamic
ZNF114	163071	no	yes	dynamic
ZNF18	7566	no	yes	dynamic
ZNF238	10472	no	yes	dynamic
ZNF306	80317	no	yes	dynamic

ZNF323	64288	no	yes	dynamic
ZNF33B	7582	no	yes	dynamic
ZNF500	26048	no	yes	dynamic
ARID3B	10620	no	yes	static
ASCL2	430	no	yes	static
ATBF1	463	no	yes	static
BAZ2B	29994	no	yes	static
BBX	56987	no	yes	static
BEX1	55859	no	yes	static
BHLHB5	27319	no	yes	static
BRD7	29117	no	yes	static
CEBPE	1053	no	yes	static
CLOCK	9575	no	yes	static
CREBL2	1389	no	yes	static
CRI2	163126	no	yes	static
DMRT1	1761	no	yes	static
E2F5	1875	no	yes	static
ELF3	1999	no	yes	static
ETV4	2118	no	yes	static
FOXD2	2306	no	yes	static
FOXD4	2298	no	yes	static
FOXD4L1	200350	no	yes	static
FOXD4L4	349334	no	yes	static
FOXG1A	2291	no	yes	static
FOXG1B	2290	no	yes	static
FOXL2	668	no	yes	static
GMEB1	10691	no	yes	static
GRHL1	29841	no	yes	static
GTF2H4	2968	no	yes	static
GTF2IRD2	84163	no	yes	static
HDAC11	79885	no	yes	static
HMG1L1	10357	no	yes	static
HMX2	3167	no	yes	static
HOXA6	3203	no	yes	static
HOXC13	3229	no	yes	static
IRX1	79192	no	yes	static
JDP2	122953	no	yes	static
KLF12	11278	no	yes	static

LEF1	51176	no	yes	static
MDS1	4197	no	yes	static
MLLT3	4300	no	yes	static
MSX1	4487	no	yes	static
NCOA1	8648	no	yes	static
NCOA6	23054	no	yes	static
NFXL1	152518	no	yes	static
NKX3-1	4824	no	yes	static
NPAS1	4861	no	yes	static
NPAS4	266743	no	yes	static
NR2E3	10002	no	yes	static
NR2F1	7025	no	yes	static
NR2F2	7026	no	yes	static
NRF1	4899	no	yes	static
NRIP1	8204	no	yes	static
OTP	23440	no	yes	static
PAX8	7849	no	yes	static
PBX4	80714	no	yes	static
PCAF	8850	no	yes	static
PCGF2	7703	no	yes	static
PHC1	1911	no	yes	static
POLR2I	5438	no	yes	static
PPP1R13L	10848	no	yes	static
PRDM8	56978	no	yes	static
RNF157	114804	no	yes	static
RRN3	54700	no	yes	static
SMARCAD1	56916	no	yes	static
SNAPC2	6618	no	yes	static
SNAPC4	6621	no	yes	static
SNAPC5	10302	no	yes	static
SOX3	6658	no	yes	static
SP110	3431	no	yes	static
SPIC	121599	no	yes	static
SUPT3H	8464	no	yes	static
TADA2L	6871	no	yes	static
TAF1L	138474	no	yes	static
TAF6L	10629	no	yes	static
TBX1	6899	no	yes	static

TBX2	6909	no	yes	static
TBX3	6926	no	yes	static
TCEAL1	9338	no	yes	static
TEAD2	8463	no	yes	static
TEAD3	7005	no	yes	static
TEAD4	7004	no	yes	static
THRAP5	10025	no	yes	static
TULP4	56995	no	yes	static
TWIST2	117581	no	yes	static
USF1	7391	no	yes	static
VENTX	27287	no	yes	static
YAF2	10138	no	yes	static
ZFPM1	161882	no	yes	static
ZNF136	7695	no	yes	static
ZNF137	7696	no	yes	static
ZNF175	7728	no	yes	static
ZNF180	7733	no	yes	static
ZNF187	7741	no	yes	static
ZNF189	7743	no	yes	static
ZNF193	7746	no	yes	static
ZNF232	7775	no	yes	static
ZNF236	7776	no	yes	static
ZNF354A	6940	no	yes	static
ZNF394	84124	no	yes	static
ZNF45	7596	no	yes	static
ZNFN1A3	22806	no	yes	static
ZNFN1A5	64376	no	yes	static
ZNHIT4	83444	no	yes	static
ZSCAN5	79149	no	yes	static
ABT1	29777	yes	no	
ARID1B	57492	yes	no	
ATF1	466	yes	no	
ATF7	11016	yes	no	
ATF7IP	55729	yes	no	
ATRX	546	yes	no	
BACH1	571	yes	no	
BACH2	60468	yes	no	
BAZ2A	11176	yes	no	

BCL3	602	yes	no	
BRPF1	7862	yes	no	
CNOT6L	246175	yes	no	
CREB5	9586	yes	no	
CRSP6	9440	yes	no	
CTBP2	1488	yes	no	
ELK3	2004	yes	no	
ELK4	2005	yes	no	
FOXK2	3607	yes	no	
FOXN4	121643	yes	no	
FOXP2	93986	yes	no	
GTF2H2	2966	yes	no	
GTF3C4	9329	yes	no	
HAND2	9464	yes	no	
HDAC10	83933	yes	no	
HES2	54626	yes	no	
HKR2	342945	yes	no	
HMG2L1	10042	yes	no	
HMGA2	8091	yes	no	
HMX3	340784	yes	no	
ISL2	64843	yes	no	
JUNB	3726	yes	no	
KIAA0415	9907	yes	no	
KLF7	8609	yes	no	
LDB1	8861	yes	no	
MAFA	389692	yes	no	
MAFK	7975	yes	no	
MAZ	4150	yes	no	
MKL2	57496	yes	no	
MLR1	254251	yes	no	
MONDOA	22877	yes	no	
MSX2	4488	yes	no	
MYBL1	4603	yes	no	
MYCPBP	10260	yes	no	
MYT1	4661	yes	no	
NCOA2	10499	yes	no	
NFATC2	4773	yes	no	
NFIB	4781	yes	no	

NPAS2	4862	yes	no	
NR1D1	9572	yes	no	
PGR	5241	yes	no	
PHF6	84295	yes	no	
POU3F1	5453	yes	no	
POU3F3	5455	yes	no	
POU6F1	5463	yes	no	
PPARA	5465	yes	no	
PRDM2	7799	yes	no	
PRRX2	51450	yes	no	
PTF1A	256297	yes	no	
RARG	5916	yes	no	
REST	5978	yes	no	
RREB1	6239	yes	no	
SIX1	6495	yes	no	
SOX1	6656	yes	no	
TAF11	6882	yes	no	
TAF3	83860	yes	no	
TAF4B	6875	yes	no	
TBN	129685	yes	no	
TBX20	57057	yes	no	
TBX21	30009	yes	no	
TBX4	9496	yes	no	
TCEB3	6924	yes	no	
TCF7L1	83439	yes	no	
TCF8	6935	yes	no	
TEAD1	7003	yes	no	
TFDP2	7029	yes	no	
TFEC	22797	yes	no	
TFPI	7035	yes	no	
TP73	7161	yes	no	
ZBTB20	26137	yes	no	
ZFP36L2	678	yes	no	
ZNF131	7690	yes	no	
ZNF138	7697	yes	no	
ZNF141	7700	yes	no	
ZNF192	7745	yes	no	
ZNF354C	30832	yes	no	

ZNF43	7594	yes	no	
ZNF70	7621	yes	no	
ZNF85	7639	yes	no	
ZNF92	168374	yes	no	
AEBP1	165	no	no	
ALF	11036	no	no	
ALX3	257	no	no	
ALX4	60529	no	no	
ARX	170302	no	no	
ASCL1	429	no	no	
ASCL3	56676	no	no	
ASCL4	121549	no	no	
ATOH1	474	no	no	
BARHL1	56751	no	no	
BARHL2	343472	no	no	
BARX2	8538	no	no	
BCL6B	255877	no	no	
BHLHB4	128408	no	no	
BHLHB8	168620	no	no	
BRCA2	675	no	no	
BRDT	676	no	no	
BTF3L2	691	no	no	
BTF3L3	692	no	no	
C10ORF121	390010	no	no	
CDX1	1044	no	no	
CDX2	1045	no	no	
CDX4	1046	no	no	
CHX10	338917	no	no	
CIITA	4261	no	no	
CITED1	4435	no	no	
CREB3L1	90993	no	no	
CREB3L3	84699	no	no	
CREM	1390	no	no	
CRX	1406	no	no	
CSEN	30818	no	no	
CTCFL	140690	no	no	
DBX1	120237	no	no	
DLX2	1746	no	no	

DLX4	1748	no	no	
DLX5	1749	no	no	
DLX6-ambiguous	1750	no	no	
DMBX1	127343	no	no	
DMRTB1	63948	no	no	
DUX1	26584	no	no	
DUX2	26583	no	no	
DUX3	26582	no	no	
DUX4	22947	no	no	
DUX5	26581	no	no	
EBF	1879	no	no	
EBF2	64641	no	no	
EHF	26298	no	no	
ELF5	2001	no	no	
EMX1	2016	no	no	
EMX2	2018	no	no	
EN1	2019	no	no	
EN2	2020	no	no	
ERG	2078	no	no	
ESR1	2099	no	no	
ESR2	2100	no	no	
ESRRB	2103	no	no	
ESRRG	2104	no	no	
ESX1L	80712	no	no	
ETV1	2115	no	no	
ETV2	2116	no	no	
ETV7	51513	no	no	
EVI1	2122	no	no	
EVX1	2128	no	no	
EVX2	2129	no	no	
FEV	54738	no	no	
FHAD1	114827	no	no	
FHL5	9457	no	no	
FKHL18	2307	no	no	
FLJ36749	283571	no	no	
FOXA1	3169	no	no	
FOXA2	3170	no	no	
FOXA3	3171	no	no	

FOXB1	27023	no	no	
FOXC1	2296	no	no	
FOXC2	2303	no	no	
FOXD3	27022	no	no	
FOXD4L2	286380	no	no	
FOXD4L3	387054	no	no	
FOXE1	2304	no	no	
FOXE3	2301	no	no	
FOXF1	2294	no	no	
FOXG1C	2292	no	no	
FOXH1	8928	no	no	
FOXI1	2299	no	no	
FOXJ1	2302	no	no	
FOXL1	2300	no	no	
FOXN1	8456	no	no	
FOXO1B	2311	no	no	
FOXO3B	2310	no	no	
FOXO6	343552	no	no	
FOXP3	50943	no	no	
FOXQ1	94234	no	no	
FOXR1	283150	no	no	
FOXR2	139628	no	no	
GATA1	2623	no	no	
GATA3	2625	no	no	
GATA4	2626	no	no	
GATA5	140628	no	no	
GATA6	2627	no	no	
GBX1	2636	no	no	
GBX2	2637	no	no	
GCM1	8521	no	no	
GCM2	9247	no	no	
GFI1B	8328	no	no	
GLI1	2735	no	no	
GLI2	2736	no	no	
GLI3	2737	no	no	
GLIS2	84662	no	no	
GRHL2	79977	no	no	
GRHL3	57822	no	no	

GRLF1	2909	no	no	
GSC	145258	no	no	
GSCL	2928	no	no	
GSH1	219409	no	no	
GSH2	170825	no	no	
GTF2F2L	2964	no	no	
HAND1	9421	no	no	
HES5	388585	no	no	
HES7	84667	no	no	
HEY1	23462	no	no	
HIC1	3090	no	no	
HIF3A	64344	no	no	
HLF	3131	no	no	
HMG17L2	23606	no	no	
HMG1L10	27126	no	no	
HMG1L3	10356	no	no	
HMG1L4	10355	no	no	
HMG1L5	10354	no	no	
HMG1L6	10353	no	no	
HMG4L	128872	no	no	
HMG4L2	128879	no	no	
HMGA1L1	203477	no	no	
HMGA1L2	171559	no	no	
HMGA1L3	144712	no	no	
HMX1	3166	no	no	
HNF4A	3172	no	no	
HOXA1	3198	no	no	
HOXA2	3199	no	no	
HOXA3	3200	no	no	
HOXA4	3201	no	no	
HOXA7	3204	no	no	
HOXB1	3211	no	no	
HOXB13	10481	no	no	
HOXB2	3212	no	no	
HOXB3	3213	no	no	
HOXB5	3215	no	no	
HOXB6	3216	no	no	
HOXB7	3217	no	no	

HOXB8	3218	no	no	
HOXB9	3219	no	no	
HOXC10	3226	no	no	
HOXC11	3227	no	no	
HOXC12	3228	no	no	
HOXC4	3221	no	no	
HOXC5	3222	no	no	
HOXC6	3223	no	no	
HOXC8	3224	no	no	
HOXC9	3225	no	no	
HOXD1	3231	no	no	
HOXD10	3236	no	no	
HOXD11	3237	no	no	
HOXD12	3238	no	no	
HOXD13	3239	no	no	
HOXD3	3232	no	no	
HOXD4	3233	no	no	
HOXD8	3234	no	no	
HOXD9	3235	no	no	
HSF4	3299	no	no	
HSFY1	86614	no	no	
HSFY2	159119	no	no	
ID2B	84099	no	no	
ID4	3400	no	no	
INSAF	3637	no	no	
IPF1	3651	no	no	
IRF4	3662	no	no	
IRF6	3664	no	no	
IRX2	153572	no	no	
IRX4	50805	no	no	
IRX6	79190	no	no	
ISL1	3670	no	no	
KLF1	10661	no	no	
KLF14	136259	no	no	
KLF15	28999	no	no	
KLF3	51274	no	no	
KLF5	688	no	no	
KLF8	11279	no	no	

L3MBTL	26013	no	no	
LASS3	204219	no	no	
LBX1	10660	no	no	
LDB2	9079	no	no	
LHX1	3975	no	no	
LHX3	8022	no	no	
LHX4	89884	no	no	
LHX5	64211	no	no	
LHX6	26468	no	no	
LHX8	431707	no	no	
LHX9	56956	no	no	
LISCH7	51599	no	no	
LMO1	4004	no	no	
LMO3	55885	no	no	
LMX1A	4009	no	no	
LMX1B	4010	no	no	
LOC202201	202201	no	no	
LOC257468	257468	no	no	
LOC285563	285563	no	no	
LOC285697	285697	no	no	
LOC340260	340260	no	no	
LOC340765	340765	no	no	
LOC342900	342900	no	no	
LOC344191	344191	no	no	
LOC360030	360030	no	no	
LOC390259	390259	no	no	
LOC390338	390338	no	no	
LOC390874	390874	no	no	
LOC391742	391742	no	no	
LOC391745	391745	no	no	
LOC391746	391746	no	no	
LOC391747	391747	no	no	
LOC391749	391749	no	no	
LOC391761	391761	no	no	
LOC391763	391763	no	no	
LOC391764	391764	no	no	
LOC391766	391766	no	no	
LOC392152	392152	no	no	

LOC399839	399839	no	no	
LOC401860	401860	no	no	
LOC401861	401861	no	no	
LOC402199	402199	no	no	
LOC402200	402200	no	no	
LOC402201	402201	no	no	
LOC402202	402202	no	no	
LOC402203	402203	no	no	
LOC402204	402204	no	no	
LOC402205	402205	no	no	
LOC402206	402206	no	no	
LOC402207	402207	no	no	
LOC402208	402208	no	no	
LOC402209	402209	no	no	
LOC402210	402210	no	no	
LOC402211	402211	no	no	
LOC402714	402714	no	no	
LOC94431	94431	no	no	
MBD3L1	85509	no	no	
MBD3L2	125997	no	no	
MDFI	4188	no	no	
MEIS1	4211	no	no	
MEIS3	56917	no	no	
MEOX1	4222	no	no	
MEOX2	4223	no	no	
MGC20410	116071	no	no	
MIXL1	83881	no	no	
MKRN3	7681	no	no	
MLLT1	4298	no	no	
MORF4	10934	no	no	
MYCL2	4611	no	no	
MYCLK1	4612	no	no	
MYCN	4613	no	no	
MYF5	4617	no	no	
MYF6	4618	no	no	
MYOCD	93649	no	no	
MYOD1	4654	no	no	
MYOG	4656	no	no	

MYT1L	23040	no	no	
MYT2	8827	no	no	
NANOG	79923	no	no	
NANOGP8	388112	no	no	
NEUROD1	4760	no	no	
NEUROD2	4761	no	no	
NEUROG1	4762	no	no	
NEUROG2	63973	no	no	
NEUROG3	50674	no	no	
NFATC4	4776	no	no	
NFE2L3	9603	no	no	
NFIA	4774	no	no	
NHLH1	4807	no	no	
NHLH2	4808	no	no	
NKX1-1	54729	no	no	
NKX2-3	159296	no	no	
NKX2-4	4823	no	no	
NKX2-5	1482	no	no	
NKX2-6	137814	no	no	
NKX2-8	26257	no	no	
NKX6-1	4825	no	no	
NKX6-2	84504	no	no	
NOBOX	135935	no	no	
NOTO	344022	no	no	
NPAS3	64067	no	no	
NR0B1	190	no	no	
NR0B2	8431	no	no	
NR1H4	9971	no	no	
NR1I2	8856	no	no	
NR1I3	9970	no	no	
NR2E1	7101	no	no	
NR3C2	4306	no	no	
NR5A1	2516	no	no	
NR5A2	2494	no	no	
NR6A1	2649	no	no	
NRIP2	83714	no	no	
NRL	4901	no	no	
OLIG3	167826	no	no	

ONECUT1	3175	no	no	
OTEX	158800	no	no	
OTX2	5015	no	no	
PAWR	5074	no	no	
PAX1	5075	no	no	
PAX2	5076	no	no	
PAX3	5077	no	no	
PAX4	5078	no	no	
PAX5	5079	no	no	
PAX6	5080	no	no	
PAX7	5081	no	no	
PAX9	5083	no	no	
PBX1	5087	no	no	
PEG3	5178	no	no	
PEPP-2	84528	no	no	
PER3	8863	no	no	
PHOX2A	401	no	no	
PHOX2B	8929	no	no	
PITX1	5307	no	no	
PITX2	5308	no	no	
PITX3	5309	no	no	
PKNOX2	63876	no	no	
PLAG1	5324	no	no	
POLR3G	10622	no	no	
POU1F1	5449	no	no	
POU2AF1	5450	no	no	
POU2F3	25833	no	no	
POU3F2	5454	no	no	
POU3F4	5456	no	no	
POU4F2	5458	no	no	
POU4F3	5459	no	no	
POU5F1	5460	no	no	
POU6F2	11281	no	no	
PPARAL	5466	no	no	
PPARGC1A	10891	no	no	
PRDM11	56981	no	no	
PRDM16	63976	no	no	
PRDM5	11107	no	no	

PRDM7	11105	no	no	
PROP1	5626	no	no	
PROX1	5629	no	no	
PRRX1	5396	no	no	
PRRXL1	117065	no	no	
RABEP2	79874	no	no	
RARB	5915	no	no	
RAX	30062	no	no	
RAXLX	91464	no	no	
RBAK	57786	no	no	
RBPSUHL	11317	no	no	
RFPL1	5988	no	no	
RFPL2	10739	no	no	
RFX4	5992	no	no	
RFXDC1	222546	no	no	
RNF2	6045	no	no	
RORA	6095	no	no	
RORB	6096	no	no	
RORC	6097	no	no	
RUNX1T1	862	no	no	
RXRG	6258	no	no	
SALF	286749	no	no	
SCML1	6322	no	no	
SCML4	256380	no	no	
SCRT1	83482	no	no	
SCXA	333927	no	no	
SHOX	6473	no	no	
SIM1	6492	no	no	
SIM2	6493	no	no	
SIX2	10736	no	no	
SIX3	6496	no	no	
SIX4	51804	no	no	
SIX6	4990	no	no	
SMARCA1	6594	no	no	
SMYD1	150572	no	no	
SNAI2	6591	no	no	
SOX10	6663	no	no	
SOX11	6664	no	no	

SOX14	8403	no	no	
SOX15	6665	no	no	
SOX17	64321	no	no	
SOX2	6657	no	no	
SOX21	11166	no	no	
SOX5	6660	no	no	
SOX6	55553	no	no	
SOX7	83595	no	no	
SOX8	30812	no	no	
SOX9	6662	no	no	
SP6	80320	no	no	
SP7	121340	no	no	
SP8	221833	no	no	
SPDEF	25803	no	no	
SPIB	6689	no	no	
SRY	6736	no	no	
SSX1	6756	no	no	
SSX2	6757	no	no	
SSX4	6759	no	no	
SSX5	6758	no	no	
ST18	9705	no	no	
STAT4	6775	no	no	
SUPT4H2	6828	no	no	
T	6862	no	no	
TAF2GL	163088	no	no	
TAF7L	54457	no	no	
TAL1	6886	no	no	
TBR1	10716	no	no	
TBX10	347853	no	no	
TBX15	6913	no	no	
TBX19	9095	no	no	
TBX22	50945	no	no	
TBX5	6910	no	no	
TBX6	6911	no	no	
TCEA3	6920	no	no	
TCEB3B	51224	no	no	
TCEB3C	162699	no	no	
TCF1	6927	no	no	

TCF15	6939	no	no	
TCF2	6928	no	no	
TCF20	6942	no	no	
TCF21	6943	no	no	
TCF23	150921	no	no	
TCF7	6932	no	no	
TCF7L2	6934	no	no	
TFAP2B	7021	no	no	
TFAP2D	83741	no	no	
TFAP2E	339488	no	no	
TFCP2L1	29842	no	no	
TFDP3	51270	no	no	
TGIF2LX	90316	no	no	
TGIF2LY	90655	no	no	
THRB	7068	no	no	
TITF1	7080	no	no	
TLE2	7089	no	no	
TLX1	3195	no	no	
TLX2	3196	no	no	
TLX3	30012	no	no	
TP53	7157	no	no	
TP73L	8626	no	no	
TXK	7294	no	no	
UTF1	8433	no	no	
VAX1	11023	no	no	
VSX1	30813	no	no	
ZBTB16	7704	no	no	
ZFHX2	85446	no	no	
ZFHX4	79776	no	no	
ZFPM2	23414	no	no	
ZNF10	7556	no	no	
ZNF11A	7557	no	no	
ZNF123	7677	no	no	
ZNF125	7679	no	no	
ZNF126	7680	no	no	
ZNF132	7691	no	no	
ZNF154	7710	no	no	
ZNF157	7712	no	no	

ZNF165	7718	no	no	
ZNF167	55888	no	no	
ZNF181	339318	no	no	
ZNF19	7567	no	no	
ZNF197	10168	no	no	
ZNF206	84891	no	no	
ZNF218	128553	no	no	
ZNF287	57336	no	no	
ZNF311	282890	no	no	
ZNF320	117040	no	no	
ZNF33A	7581	no	no	
ZNF345	25850	no	no	
ZNF354B	117608	no	no	
ZNF37A	7587	no	no	
ZNF396	252884	no	no	
ZNF41	7592	no	no	
ZNF423	23090	no	no	
ZNF435	80345	no	no	
ZNF44	51710	no	no	
ZNF482	10773	no	no	
ZNF483	158399	no	no	
ZNF537	57616	no	no	
ZNF71	58491	no	no	
ZNF72	7623	no	no	
ZNF73	7624	no	no	
ZNF80	7634	no	no	
ZNF81	347344	no	no	
ZNF90	7643	no	no	
ZNF93	7646	no	no	
ZNF96	9753	no	no	
ZNFN1A2	22807	no	no	
ZNFN1A4	64375	no	no	
ZSCAN4	201516	no	no	

Supplementary Table 13 The enriched TF motifs in the promoters of TF clusters. Only top 10 motifs are shown. All data sets are available from the FANTOM4 web resource.

a) Motifs enriched in the promoters of Down-regulated TFs

MOTIF	(n) set	(n) all	p-val	%set	%all	x enrichment
GATA4	14	40	1.45E-05	22%	7%	3.3
OCT4	14	40	1.45E-05	22%	7%	3.3
NFYA,B,C	23	98	4.51E-05	36%	16%	2.2
MAZ	1	103	5.42E-05	2%	17%	0.1
GFI1B	7	14	0.000208	11%	2%	4.8
TBX5	11	34	0.000285	17%	6%	3.1
Helios	12	41	0.000418	19%	7%	2.8
GTF2I	20	89	0.000476	31%	15%	2.1
YY1	17	74	0.000731	27%	12%	2.2
MAZR/ZNF278	1	77	0.001199	2%	13%	0.1

b) Motifs enriched in the promoters of Up-regulated TFs

MOTIF	(n) set	(n) all	p-val	%set	%all	x enrichment
MAZ	25	103	1.39E-13	74%	17%	4.4
CRE-BP1:c-Jun	10	26	2.30E-07	29%	4%	6.9
FALZ	7	23	0.000112	21%	4%	5.5
SNAI1-3	8	31	0.000123	24%	5%	4.6
TBP	7	24	0.000151	21%	4%	5.2
Nkx2-2	5	14	0.000535	15%	2%	6.4
NFATC1-3	8	38	0.000549	24%	6%	3.8
ELF1/2/4	10	59	0.000637	29%	10%	3
IRF7	7	30	0.000665	21%	5%	4.2
FOXD1/2	5	15	0.000763	15%	2%	6

c) Motifs enriched in the promoters of transiently regulated TFs

MOTIF	(n) set	ALL (n)	p-val	%set	%all	x enrichment
SRF	7	12	0.001169	7%	2%	3.5
NHLH1/2	7	14	0.003774	7%	2%	3
ELK1/4_GABPA/B2	7	94	0.004107	7%	15%	0.4
GFI1	12	33	0.005039	12%	5%	2.2

IRF1,2	7	16	0.007473	7%	3%	2.6
IRF7	11	30	0.007676	11%	5%	2.2
FOSL2	8	22	0.014177	8%	4%	2.2
AREB6	7	18	0.014659	7%	3%	2.3
EBF1	5	11	0.019258	5%	2%	2.7
NRF1	8	89	0.019331	8%	15%	0.5

d) Motifs enriched in the promoters of TFs induced in the first hour

MOTIF	(n) set	(n) all	p-val	%set	%all	x enrichment
AGL3	6	15	5.23E-07	38%	2%	15.3
SRF	5	12	4.57E-06	31%	2%	15.9
FOSL2	5	22	0.000126	31%	4%	8.7
Agamous	5	27	0.000351	31%	4%	7.1
TBP	4	24	0.002309	25%	4%	6.4
SP1	0	190	0.002329	0%	31%	0
CRE-BP1:c-Jun	4	26	0.003115	25%	4%	5.9
SOX8-10	4	28	0.004092	25%	5%	5.4
PBX1	4	30	0.005251	25%	5%	5.1
CRE-BP1	3	19	0.010283	19%	3%	6

Supplementary Table 14 Frequency of leukemia related terms in entrez gene annotations for transcription factors

a) Gene counts:

Transcription factor class	cancer	leukemia	myeloid leuk	lymphoma	ALL
Transient	38	23	7	11	101
Undifferentiated	24	17	10	4	64
Differentiated	13	5	0	5	34
Static	111	60	25	20	411
not detected	136	59	18	27	712
all TFs	322	164	60	67	1322
detected	186	105	42	40	610

b) Percentages:

Transcription factor class	cancer	leukemia	myeloid leuk	lymphoma
Transient	38%	23%	7%	11%
Undifferentiated	38%	27%	16%	6%
Differentiated	38%	15%	0%	15%
Static	27%	15%	6%	5%
not detected	19%	8%	3%	4%
all TFs	24%	12%	5%	5%
detected	30%	17%	7%	7%

c) p-val: (compared to all TFs)

Transcription factor class	cancer	leukemia	myeloid leuk	lymphoma
Transient	2.66E-03	1.84E-03	1.34E-01	1.11E-02
Undifferentiated	1.80E-02	1.10E-03	3.27E-04	3.40E-01
Differentiated	4.68E-02	3.62E-01	2.02E-01	1.98E-02
Static	4.58E-01	1.43E-01	8.22E-02	4.36E-01

Frequency of leukemia related terms in Entrez gene annotations for transcription factors down-regulated, up-regulated and transiently induced/repressed during PMA-induced differentiation. (a) number of TFs from each class with terms cancer, leukemia, ‘myeloid leukemia’, and lymphoma. (b) percentage of TFs from each class with these terms. (c) p-value for the observation using fisher's exact test (background used is all TFs).

Supplementary Table 15 Accession numbers of the data set in the public database.

Category and Database	Data	Accession No.
CAGE, DDBJ	THP-1 PMA stimulated 6 time points (0,1,4,12,24,96 hours) (RIKEN1,RIKEN3,RIKEN6)	rna_lib_id definition accession No. HFH THP-1 PMA 1h RIKEN1 AFAAA0000001-AFAAA0543260 (543260 entries) HFI THP-1 PMA 4h RIKEN1 AFAAB0000001-AFAAB0191814 (191814 entries) HFJ THP-1 PMA 12h RIKEN1 AFAAC0000001-AFAAC0474745 (474745 entries) HFK THP-1 PMA 24h RIKEN1 AFAAD0000001-AFAAD0353461 (353461 entries) HGC THP-1 PMA 96h RIKEN1 AFAAE0000001-AFAAE0270208 (270208 entries) HGS THP-1 PMA 0h RIKEN1 AFAAF0000001-AFAAF0513982 (513982 entries) HHA THP-1 PMA 0h RIKEN3 AFAAG0000001-AFAAG0349448 (349448 entries) HHB THP-1 PMA 1h RIKEN3 AFAAH0000001-AFAAH0425524 (425524 entries) HHC THP-1 PMA 4h RIKEN3 AFAAI0000001-AFAAI0358861 (358861 entries) HHD THP-1 PMA 12h RIKEN3 AFAAJ0000001-AFAAJ0285303 (285303 entries) HHE THP-1 PMA 24h RIKEN3 AFAAK0000001-AFAAK0327328 (327328 entries) HHF THP-1 PMA 96h RIKEN3 AFAAL0000001-AFAAL0456381 (456381 entries) HHG THP-1 PMA 0h RIKEN6 AFAAM0000001-AFAAM0436519 (436519 entries) HHH THP-1 PMA 1h RIKEN6 AFAAN0000001-AFAAN0499442 (499442 entries) HHI THP-1 PMA 4h RIKEN6 AFAAO0000001-AFAAO0682996 (682996 entries)
Array, CIBEX	Affymetrix whole genome array THP-1 PMA stimulated 2 time points (0,96 hours) H3K9 acetylation	CBX48
	Affymetrix whole genome array THP-1 PMA stimulated 2 time points (0,96 hours) Pol2 stimulated data	CBX44
	Affymetrix promoter array THP-1 PMA stimulated 2 time points (0,96 hours) PU.1 stimulated data	CBX43
	Affymetrix promoter array THP-1 PMA stimulated 2 time points (0,96 hours) SP1 stimulated data	

	Illumina THP-1 PMA stimulated 10 time points (0,1,2,4,6,12,24,48,72,96 hours) 3	CBX46
	Illumina THP-1 54 genes knocked down data 3 replica	CBX47

CIBEX (Center for Information Biology gene Expression database)
DDBJ (DNA Data Bank of Japan)